The Map Becomes the Model: A Taxonomy of Worldwide AI Fragmentation by Topic, Jurisdiction, and Owner Policy

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

December 27, 2025

AI systems are often discussed as if they were universal instruments: one set of weights, one interface, one "truth engine," deployable anywhere a network exists. In practice, AI is becoming geographically conditional. The moment a system answers questions about weapons, drugs, sex, elections, dissent, surveillance, or medical and legal risk, it collides with jurisdictional boundaries that do not align. Nations diverge on speech law, security doctrine, privacy expectations, and moral thresholds; even within a single country, state and provincial statutes can invert permissions over a short drive. On top of law sits the provider's own risk posture: global safety floors, liability avoidance, brand constraints, and selective market access. The result is not merely "bias" in the usual sense, but a structural phenomenon: the same underlying capability produces different allowable outputs depending on location, topic, intent, and product tier. This paper proposes a practical taxonomy of AI fragmentation. It maps high-fragmentation topics, identifies the governing forces behind divergence, and describes predictable failure modes such as overblocking, self-sealing refusals, and jurisdictional misrouting. Finally, it sketches a design approach for "jurisdiction-aware AI" that surfaces which layer made a decision, supports contestability, and avoids the illusion that one model can speak everywhere in the same voice.

Keywords: AI governance, jurisdiction, policy overlays, content moderation, geofencing, safety floor, compliance packs, sovereignty, censorship, liability, dual use, speech law, privacy regimes, election integrity, weapons policy, controlled substances, sexual content regulation, hate speech, extremism, surveillance, medical advice, legal advice, sanctions, platform incentives, auditability, transparency, contestability, model routing.

A collaboration with GPT-5.2-Thinking. 57 pages. CC4.0.

Outline

1. Introduction: When the Map Enters the Interface
1.1 The core claim: AI capability is global, but AI permission is local
1.2 Why “one model worldwide” breaks on high-stakes topics
1.3 From “bias debates” to “fragmentation mechanics”
1.4 Definitions: fragmentation, jurisdiction pack, safety floor, owner policy

2. The Fragmentation Stack: Three Sovereignties and One Output
2.1 State sovereignty: law, security doctrine, and enforcement incentives
2.2 Owner sovereignty: liability, brand, market access, and internal ideology
2.3 Model sovereignty: training priors, alignment rules, and refusal styles
2.4 The strictest-layer effect: why the harshest constraint dominates

3. A Taxonomy of High-Fragmentation Topics
3.1 Weapons, self-defense, explosives, and operational violence
3.2 Drugs and controlled substances: state-federal contradictions and thresholds
3.3 Sex, pornography, sex work, and age-gated speech
3.4 LGBTQ speech and identity: legality, “propaganda,” and criminalization risk
3.5 Hate speech, group vilification, and extremist propaganda
3.6 Terrorism, insurgency, and “material support” boundaries
3.7 Elections, persuasion, and targeted political messaging
3.8 Dissent, protest logistics, and criticism of state institutions
3.9 Surveillance, biometrics, and policing technologies
3.10 Privacy, doxxing, and re-identification through aggregation
3.11 Cybersecurity, malware, and dual-use technical detail
3.12 Medical advice, reproductive health, and controlled medications
3.13 Legal advice: firearms/drugs transport, immigration, criminal exposure
3.14 Finance: crypto, gambling, sanctions, and evasion-adjacent requests
3.15 Religion and blasphemy constraints in specific jurisdictions

4. Drivers of Divergence: Why the Same Prompt Breaks Differently
4.1 Criminal law variance: definitions, thresholds, and intent standards
4.2 Speech law variance: illegal content categories and duty-of-care regimes
4.3 Security variance: “state stability” doctrines and forbidden narratives
4.4 Privacy variance: data localization, sensitive data, and identification rules
4.5 Civil liability variance: negligence, defamation, and product safety exposure
4.6 Cultural variance: social enforcement, not just legal enforcement

5. Failure Modes: What Fragmentation Looks Like to Users
5.1 Overblocking: safe information refused because routing is too blunt
5.2 Underblocking: harmful operational detail slips through in “neutral” framing
5.3 Self-sealing refusals: non-falsifiable policy language that escalates distrust
5.4 Jurisdictional misrouting: wrong legal regime applied to the user

5.5 Asymmetric enforcement: perceived favoritism caused by unequal constraint sets
5.6 The whiplash effect: crossing a border changes the model's persona

6. Measuring Fragmentation: Metrics that Replace Vibes with Evidence
6.1 Topic volatility index: how often rules differ across jurisdictions
6.2 Refusal gradient mapping: where refusals cluster and why
6.3 Parity tests: symmetrical handling of comparable claims about different groups
6.4 Compliance drift tracking: when policies change silently over time
6.5 User trust metrics: contestability, clarity, and appeal success rates

7. Design Patterns for Jurisdiction-Aware AI
7.1 The global safety floor: invariants that should not vary
7.2 Jurisdiction packs: modular legal overlays with explicit provenance
7.3 Use-case packs: enterprise, education, clinical, newsroom, research modes
7.4 Policy routers: intent detection, location detection, and conservative fallbacks
7.5 Explanation interfaces: "which layer constrained this answer"
7.6 Audit logs: reproducibility, versioning, and policy changelogs

8. Governance: Avoiding Oracle Politics While Staying Deployable
8.1 Contestability: appeals, reversals, and reasons that can be challenged
8.2 Third-party audits: eval harnesses that cross jurisdictions
8.3 Forkability and exit: reducing monopoly control over "allowable speech"
8.4 Transparency without leakage: how to explain constraints safely
8.5 The non-universal conclusion: plural systems and the end of one-size-fits-all

9. Case Studies: The Border as a Demonstration Tool
9.1 US state borders: firearms, cannabis, and licensing thresholds
9.2 EU compliance regimes vs. US liability regimes: different refusal styles
9.3 China/Russia: sovereignty-first approaches and external model incompatibility
9.4 Corporate policy conflicts: reputation risk vs. user autonomy

10. Conclusion: The Map Becomes a First-Class Parameter
10.1 Thesis restated: capability global, permission local
10.2 What is gained by admitting fragmentation: honesty, auditability, trust
10.3 A forward path: modularity, disclosure, and contestable governance

References

Art

1. Introduction: When the Map Enters the Interface

For a brief window, it was possible to talk about AI systems as if they were simply "on the internet," and therefore universally available and universally coherent. A model could be trained once, deployed once, and then "the world" could ask it questions. That mental picture is already obsolete. The frontier issue is not whether AI can answer questions, but whether it is allowed to answer them, how it must answer them, and who decides what "allowed" means in each place. The map is no longer a background condition of deployment. It is entering the interface as an active parameter that shapes the content, the tone, the refusal boundaries, and even the confidence posture of the system.

The shift matters because it changes what users are actually interacting with. A user thinks they are interacting with "the model," but in practice they are interacting with an output assembly line: base capability plus policy filters plus jurisdiction-specific compliance logic plus company risk management plus runtime routing decisions. When this stack behaves differently in different places, users will naturally reach for moral explanations ("bias," "agenda," "propaganda," "censorship") because the behavior feels personal. Yet the deeper pattern is structural. The same capability is being governed by multiple sovereigns at once: states, owners, and the designers of the safety regime.

This paper starts with a simple claim that reframes the debate: AI capability is increasingly global, but AI permission is irreducibly local. Once this is granted, the question becomes: what topics fragment first, what mechanisms cause the fragmentation, what failure modes follow, and what design patterns can preserve usefulness without pretending that a single model can speak everywhere with the same voice.

1.1 The core claim: AI capability is global, but AI permission is local

Modern foundation models are trained on large, cross-border corpora, optimized on hardware that is globally traded, and deployed on networks that ignore national boundaries at the packet level. This produces an intuitive belief that "capability" is placeless. If the weights exist, they exist everywhere. If the model can solve a problem, it can solve it for anyone.

But permission is not placeless. Permission is the layer where law, norms, institutional incentives, and enforcement realities converge. Permission is the difference between "can the model answer this?" and "may the system provide this answer here, to this user, in this context, in this form?" Permission includes what must be refused, what must be phrased cautiously, what must be age-gated, what must be logged, what must be escalated to a human, what must be region-blocked, and what must be reported.

The consequence is that a user's lived experience of AI will increasingly be determined by location and jurisdiction. Two users may type the same prompt and encounter different refusals,

different hedges, different safety explanations, different resource suggestions, and different degrees of detail. This is not a bug in the strict sense; it is a predictable outcome of deploying powerful language systems into a world where legal and moral regimes remain plural.

So the paper's core claim is a reframing: the global spread of capability does not imply a global spread of permission. In fact, as capability increases, permission constraints often tighten, because the risk surface grows with competence. The map enters the interface precisely because the model becomes useful enough to matter.

1.2 Why "one model worldwide" breaks on high-stakes topics

Low-stakes topics tolerate uniformity. If a user asks about weather metaphors, chess openings, or the history of steam engines, jurisdictions rarely demand different answers. But high-stakes topics are exactly where governments, platforms, and civil society insist on different rules. "High-stakes" here means any topic that reliably touches one or more of these triggers:

First, criminal exposure. Topics like weapons, controlled substances, explosives, hacking, or fraud are not merely cultural disputes; they are areas where detailed instructions can create immediate legal consequences. In one jurisdiction, a behavior is permitted and normalized; in another, it is prosecuted aggressively; in a third, it is politically instrumentalized. A model that gives operational detail in the wrong place can be blamed for enabling harm, and the provider can be blamed for distributing an unlicensed advisor.

Second, political stability and regime legitimacy. Elections, protest logistics, criticism of leaders, secession narratives, or state-security controversies draw hard lines. Some regimes treat certain speech acts as threats to sovereignty or public order. The "one model worldwide" fantasy fails here because different states are not merely disagreeing about facts; they are disagreeing about what kinds of speech are legally tolerable.

Third, moral boundary enforcement. Sex, pornography, sex work, and identity-related speech vary radically across cultures and legal systems, including strong divergence on what must be age-gated, restricted, or criminalized. A model that behaves uniformly will be illegal in some places, morally condemned in others, and competitively displaced by localized alternatives that match local expectations.

Fourth, privacy and surveillance. Some regimes demand strong privacy protections and limit biometric uses; others explicitly prioritize public security and administrative control. This yields opposite requirements for logging, identification, retention, and compliance reporting. One model cannot satisfy both simultaneously without becoming effectively multiple systems.

Fifth, the owner's liability surface. Even if the law is permissive, the provider may still refuse to reduce legal risk, reputational risk, or misuse risk. This is the corporate "permission layer," and it

is global to the platform even when the law is local. A user sees this as “the AI won’t answer,” but the deeper reason is that the provider cannot accept the downside of being too helpful on a topic that can be weaponized.

High-stakes topics therefore function as fracture planes. They reveal where political pluralism, legal pluralism, and corporate risk management collide. The more powerful the system, the more those collisions matter, and the less plausible it becomes that a single uniform conversational agent can be “for everyone everywhere.”

1.3 From “bias debates” to “fragmentation mechanics”

In public discourse, differences in AI behavior are often diagnosed as “bias.” That diagnosis is sometimes correct in the narrow sense: models can reflect skewed training data, unequal error rates, and inconsistent application of moderation rules. But the bias framing tends to personalize and moralize what is often a structural phenomenon. It implies the model has a single worldview that it is imposing.

Fragmentation mechanics offers a more useful lens. Instead of asking “which ideology is the model pushing,” it asks: what constraints are applied, at what layer, by whom, and under what incentives? The relevant questions become operational:

- Which jurisdiction is assumed, and how is jurisdiction detected or inferred?
- Which policy profile is active, and what triggers escalate constraints?
- Is the refusal driven by a global safety floor, a local law overlay, a product tier rule, or a corporate risk posture?
- Are different user segments routed to different versions or tools?
- Do policy changes occur silently over time, and how does that affect reproducibility?

When we shift from bias debates to fragmentation mechanics, the goal changes. The goal is no longer to prove neutrality, which is both impossible and strategically naive. The goal is to make the system legible: to make outputs attributable to layers, to reduce inconsistent behavior that looks arbitrary, and to preserve contestability.

This shift also defuses a common failure mode: the user assumes that any refusal is a moral condemnation, or that any guardrail is evidence of hidden political control. In reality, many refusals are mundane: legal risk, misuse risk, or compliance requirements. The fragmentation lens does not deny politics; it simply places politics inside the governing stack, where it can be named and audited rather than smuggled in through tone and implication.

1.4 Definitions: fragmentation, jurisdiction pack, safety floor, owner policy

To keep the rest of the paper precise, the following terms are used consistently.

Fragmentation: The systematic divergence of AI outputs, refusal boundaries, detail levels, or recommended actions across jurisdictions, user contexts, or product tiers, given prompts that are substantially similar in intent. Fragmentation is not merely random inconsistency; it is a patterned result of differing constraints.

Jurisdiction pack: A modular set of rules and transformations applied at runtime to align the system's behavior with a specific legal and regulatory environment. A jurisdiction pack may include topic restrictions, required warnings, age gating, logging mandates, data retention constraints, mandated escalation pathways, and locale-specific defamation or privacy handling. A jurisdiction pack may be applied based on geolocation, account region, enterprise compliance settings, or explicit user declaration. Crucially, a jurisdiction pack does not change the underlying capability in principle; it changes the permissions and presentation.

Safety floor: A baseline set of prohibitions and conservative behaviors that remain invariant across jurisdictions, intended to prevent the system from enabling severe harms. The safety floor is the provider's attempt to define non-negotiable constraints that apply everywhere regardless of local permissiveness. In practice, the safety floor often includes constraints related to violence facilitation, weapons operationalization, exploitation, certain categories of illicit behavior, and privacy violations. The key property of a safety floor is invariance: it is supposed to travel with the model.

Owner policy: The set of constraints and preferences imposed by the entity operating the system, including liability management, reputational considerations, business incentives, partnerships, and internal values. Owner policy may be stricter than local law and may override local permissiveness. Owner policy is also what tends to shift most frequently, because it responds to incidents, press cycles, regulator signals, and competitive positioning.

A crucial relationship follows from these definitions: fragmentation arises when jurisdiction packs differ, when owner policy differs by market or tier, and when the safety floor interacts with both in inconsistent ways. The central design challenge is not to eliminate fragmentation, which is unrealistic, but to manage it: to make it predictable, attributable, contestable, and auditable, so that users can distinguish "the model cannot do this" from "this deployment is not permitted to do this here."

With these definitions in place, the rest of the paper can move from diagnosis to taxonomy: which topics fragment first, what drivers produce the divergence, and what architectures can keep an AI system useful without requiring a false universalism.

2. The Fragmentation Stack: Three Sovereignties and One Output

A user experiences an AI response as a single utterance. But the utterance is the end product of a stack. That stack is governed by multiple sovereignties that do not naturally harmonize. The map enters the interface because these sovereignties each claim authority over what may be said, to whom, under what conditions, and with what consequences.

This section names the governing layers and explains why "disagreements" about AI behavior are often not disagreements about truth at all. They are disagreements about jurisdiction, risk, and power. The key diagnostic move is to stop treating the AI output as if it were the direct expression of one unified mind. In real deployments, outputs are assembled by routing logic, policy gates, safety filters, and product constraints that belong to different sovereigns. Each sovereign brings its own incentives and failure modes.

To make this concrete, imagine the response pipeline as a multi-stage decision process. A prompt arrives. The system must infer intent, detect risky content, apply policy rules, possibly consult tools, possibly refuse, possibly transform the style, and log the interaction. At each stage, different actors claim the right to define the acceptable action. Fragmentation is what happens when these claims conflict or when the system resolves them differently across places and products.

The paper's "three sovereignties" framework is not a metaphor. It is a deployment reality: state sovereignty, owner sovereignty, and model sovereignty. Together they form a "fragmentation stack" that determines what users can do with AI in practice.

2.1 State sovereignty: law, security doctrine, and enforcement incentives

State sovereignty is the most obvious source of fragmentation because it is the only one backed by coercive power. States define what is illegal, what is regulated, and what is permitted. But "law" is not just a set of rules. It is also doctrine (what the state believes is dangerous) and enforcement incentives (what it chooses to prosecute, tolerate, or selectively punish).

Law is the formal layer. It includes criminal statutes, speech restrictions, privacy requirements, consumer protection rules, licensing boundaries, and data governance obligations. These rules differ sharply across jurisdictions, and in many places they diverge not only in threshold but in kind. One state's "protected speech" is another state's "extremism." One jurisdiction's "medical advice" is another's "unlicensed practice." One place's "privacy right" is another's "security obstruction." The model may be capable of discussing all of these, but the deployment cannot ignore the jurisdiction that enforces consequences.

Security doctrine is the informal but decisive layer. Every state has categories of speech and action it treats as threats to stability. Some states encode those categories explicitly in law,

others enforce them through administrative pressure, platform licensing, and selective prosecution. Security doctrine is where politically sensitive topics live: elections, protest coordination, criticism of leaders, separatism, “rumors,” national humiliation narratives, and historical controversies. If you want to predict fragmentation, you do not start with “what is true.” You start with “what is high-stakes for the regime.”

Enforcement incentives are where the law becomes real. A statute can exist for decades without routine enforcement; then an incident happens and enforcement becomes aggressive. Enforcement incentives shift with politics, media cycles, and institutional rivalries. They also shift with resource constraints: police budgets, court capacity, and prosecutorial priorities. The practical meaning of a rule is therefore the expected consequence of violating it, not its textual definition. Two jurisdictions can have similar text but different real-world enforcement. This produces a subtle kind of fragmentation: the AI provider may comply in one place because enforcement is predictable, and be more cautious in another because enforcement is arbitrary.

State sovereignty also produces a particular demand: controllability. Many governments want AI systems to be auditable, steerable, and subordinated to local authority. This can include demands for local data retention, local model hosting, local safety review, local takedown responsiveness, and local content constraints. Even when a state does not explicitly say “we want controllability,” its regulatory posture implies it. The less controllable a foreign model appears, the more likely it is to be restricted, displaced by domestic alternatives, or treated as a security risk.

So state sovereignty fragments AI not only by banning certain outputs, but by demanding the infrastructure of compliance: logging, moderation, reporting, localization, and governance visibility. The map enters the interface because the state demands that the interface obey the map.

2.2 Owner sovereignty: liability, brand, market access, and internal ideology

Owner sovereignty is the set of constraints imposed by the company or institution operating the model. It is “sovereignty” because it is a real governing force: owners decide what the system will do, what it will refuse, what it will log, what it will prioritize, and where it will be offered. Users often treat this as a subset of “the model,” but it is distinct. The owner’s policies can change overnight without the weights changing at all.

Liability is the first driver. Owners have to anticipate lawsuits, regulatory penalties, and claims of negligence or facilitation. If a system provides operational guidance that leads to harm, the owner may be accused of enabling that harm. Even if the owner ultimately wins in court, the cost of litigation can be existential. This makes owners cautious in exactly the domains where users most desire specificity: weapons, drugs, self-harm, hacking, high-risk medical guidance,

and illegal activity. Liability does not require the owner to be morally opposed to something; it only requires the owner to be rational about downside.

Brand is the second driver. Owners respond to reputational risk: news stories, social media outrage, activist campaigns, and competitor attacks. Brand risk is not stable across cultures. A refusal that looks "responsible" in one market can look "politically captured" in another. A willingness to discuss certain topics can be viewed as "freedom" in one place and "degeneracy" or "subversion" in another. To protect brand, owners often adopt global policies that avoid the sharpest edges. This produces a predictable complaint: the system becomes "lowest-common-denominator" in public deployments, while specialized access is reserved for controlled environments.

Market access is the third driver. Owners want to operate in as many markets as possible, but each market can demand different compliance behaviors. Some demands are straightforward: data handling, consumer protections, content restrictions. Others are incompatible with the owner's global stance or with other markets' expectations. The result is market segmentation: different features, different policy profiles, different access gates, or total withdrawal. Fragmentation becomes an economic strategy as much as a governance constraint.

Internal ideology is the fourth driver and the most contested. Every organization has values and assumptions about what constitutes harm, dignity, fairness, and acceptable speech. These assumptions shape policy even when not explicitly named. The term "ideology" here is not a slur; it is a recognition that no owner is value-neutral. A provider may treat certain types of speech as inherently dangerous; another may treat them as legitimate political discourse. A provider may treat certain identity topics as protected; another may treat them as taboo. These differences often surface as inconsistent moderation patterns that users experience as selective enforcement.

Owner sovereignty therefore creates fragmentation even within a single jurisdiction. Two models in the same country can behave differently because their owners have different risk tolerances and values. And owner sovereignty can override local permissiveness: the law may allow a response, but the owner refuses to provide it anyway.

The key point is that owner sovereignty is not simply "compliance." It is strategic governance. Owners operate the interface, and therefore they control what the user perceives as "what AI is."

2.3 Model sovereignty: training priors, alignment rules, and refusal styles

Model sovereignty is the set of behaviors that arise from the model as a trained system, plus the alignment and tuning that shape its interaction patterns. It is "sovereignty" because it is not perfectly reducible to either law or owner preference. Even if you specify policies, the model

still has a learned distribution over responses, a learned notion of what kinds of language typically follow which prompts, and a learned tendency to hedge, refuse, moralize, summarize, or deflect.

Training priors matter because models absorb patterns from their training data. They learn how topics are usually discussed, what frames tend to be present, what rhetorical moves are common, and what associations link words to other words. Even if you later impose rules, these priors influence the "default" completion before the policy layer intervenes. That affects tone and framing: what the model treats as normal, what it treats as controversial, what it treats as dangerous, and what it treats as worthy of caution.

Alignment rules matter because models are tuned to follow instructions, avoid certain harms, and adopt a particular "assistant persona." This tuning shapes not only what is refused but how it is refused. Two systems can refuse the same request in radically different styles: one clinical and minimal, one moralizing, one apologetic, one accusatory, one overly verbose, one too terse. Users interpret refusal style as ideology. Often they are reacting to style more than to substance.

Refusal styles are especially important for fragmentation because they create secondary conflicts. Even when a refusal is legitimately required (by safety floor or law), the refusal can trigger distrust if it feels arbitrary, patronizing, or inconsistent with how other topics are handled. In other words, the refusal is not just a "no." It is a communicative act that can either preserve the user's dignity or provoke escalation.

Model sovereignty also includes epistemic behaviors: how the model expresses uncertainty, how it signals confidence, and whether it distinguishes facts from conjecture. In politically charged environments, epistemic posture becomes political. A model that hedges too much can be accused of evasion. A model that speaks too confidently can be accused of propaganda. A model that "balances both sides" can be accused of false equivalence. These are not trivial concerns because they determine whether the model can be trusted as an instrument of inquiry.

Finally, model sovereignty is what creates "residual fragmentation" even after you standardize policies. You can impose the same policy pack in two regions and still see differences due to language, cultural framing, and the model's learned associations. This is why policy alone cannot guarantee uniformity. The model's learned distribution remains a governing force.

2.4 The strictest-layer effect: why the harshest constraint dominates

In a multi-sovereign stack, conflicts are resolved by the most risk-averse layer. This is the strictest-layer effect: when constraints conflict, the system behaves as if the strictest applicable

constraint is the true law of reality. Users often find this maddening because it produces conservative refusals even where the user's local context is permissive.

The strictest-layer effect arises from asymmetric downside. If the system is too permissive, the downside can be severe: criminal penalties, platform bans, regulatory sanctions, lawsuits, or reputational collapse after a headline incident. If the system is too restrictive, the downside is usually softer: user frustration, lost sales, complaints about bias, migration to competitors. In risk management terms, owners prefer the latter.

This effect is amplified by uncertainty. If the system is not perfectly sure of the jurisdiction, the user's age, the user's intent, or the legality of an action, it will default to conservative assumptions. Jurisdiction detection fails at borders. Users travel. VPNs exist. Accounts are shared. Language can be ambiguous. Intent can be masked. The system cannot safely bet on permissiveness. So it declines, or it responds at a high level without operational detail.

The strictest-layer effect also emerges when the provider wants global scalability. A single global policy that is acceptable everywhere is often stricter than any one jurisdiction. This creates the "lowest-common-denominator assistant" phenomenon: safe, bland, often frustratingly general. Meanwhile, more permissive or more specialized behavior is moved behind enterprise agreements, professional licensing workflows, or regulated gateways where the provider can offload risk onto institutions.

There is also a strategic strictest-layer effect: owners sometimes adopt the strictest plausible interpretation of a rule to preempt accusations. If critics can frame an output as harmful, the owner may tighten policy beyond legal necessity to avoid future controversy. This turns public perception into a de facto regulator.

The result is a recognizable user experience: "The AI won't answer, even though it's legal here." That experience is real, but it is not best explained as moral judgment. It is best explained as the strictest-layer effect operating across three sovereignties: state law in some places, owner risk posture globally, and model alignment conservatism as a default.

The design implication is crucial. If you want jurisdiction-aware AI that remains useful, you cannot merely add more constraints. You must make constraints legible. You must tell the user which layer is acting, why the system is taking the conservative path, and what kinds of safe alternatives exist (high-level explanation, pointers to official sources, or "consult a licensed professional" gates). Without this legibility, strictness looks like arbitrary power, and arbitrary power invites backlash.

This section therefore sets up the paper's broader thesis: fragmentation is not a moral flaw in a model; it is a structural outcome of stacking sovereignties under asymmetric risk. The question

is whether we can design systems that admit this structure, surface it honestly, and preserve inquiry without pretending to be a universal oracle.

3. A Taxonomy of High-Fragmentation Topics

Not all topics fragment equally. Some content areas remain largely stable across borders because they rarely intersect criminal law, regime legitimacy, or culturally enforced taboos. High-fragmentation topics are different: they are precisely the domains where a competent model can create real-world consequences, where speech is treated as action, and where "helpfulness" can be reinterpreted as facilitation. These topics are also where policy differences are not merely about taste; they are about sovereignty, liability, and moral boundary enforcement.

This taxonomy is intended as a deployment tool. Each topic category below is presented in terms of why it fractures globally, what kinds of constraints tend to appear, and what failure modes are predictable. The point is not to moralize the categories. The point is to name the fracture planes so that a future "jurisdiction-aware" system can route, disclose, and audit its behavior rather than behaving like an opaque oracle whose refusals feel arbitrary.

3.1 Weapons, self-defense, explosives, and operational violence

This category fractures because it sits at the intersection of criminal facilitation risk and political legitimacy. The same object can be treated as ordinary property in one jurisdiction and as contraband in another. The same act can be celebrated as self-reliance in one place and framed as public endangerment in another. Beyond legality, providers also impose global safety floors because operational guidance for violence has immediate misuse potential.

Fragmentation here typically splits along three axes. First, weapon legality (what exists, who may possess it, and where). Second, use doctrine (self-defense standards, proportionality, duty to retreat, castle doctrine). Third, instruction sensitivity (how-to detail, sourcing, modification, concealment, evasion of detection). A model that is "helpful" in the wrong way becomes a liability magnet. A model that is too cautious becomes useless to users who want compliance information for lawful carry, hunting, sport, or safety.

Failure_toggle patterns: overblocking of benign safety guidance, refusal contagion where any mention of weapons triggers a block, and jurisdictional whiplash where a user crossing a border experiences sudden refusal changes.

3.2 Drugs and controlled substances: state-federal contradictions and thresholds

Drugs fragment because legality is layered. In many countries, drugs law is centralized. In others, it is federal with divergent state rules, and the contradictions are not abstract. They are enforced through traffic stops, border checks, airport rules, and employment screening. "Thresholds" matter: small possession vs distribution; personal use vs intent to sell; medical authorization vs recreational use; decriminalized vs legalized; and the presence of paraphernalia.

A major driver is that users often ask for practical guidance: "What can I do legally here," "How risky is transport," "What are the penalties," "How to pass a test," "How to dose safely," "Where to buy." Some of those are legal compliance questions, some are health questions, some are requests for evasion, and the same sentence can straddle categories. Providers tend to enforce strict global rules against facilitating illegal procurement or evasion, while allowing high-level harm reduction or legal summaries in certain contexts. That creates fragmentation: two people asking about "the same thing" may be differently classified depending on phrasing and inferred intent.

Failure modes include: misclassification of harm-reduction as facilitation, refusal styles that feel moralizing, and unsafe "partial answers" that omit key context because the model is trying to be helpful without crossing a line.

3.3 Sex, pornography, sex work, and age-gated speech

This is a high-fragmentation domain because it compresses three kinds of divergence into one: legality, moral norms, and child-safety constraints. "Sexual content" is not one thing. It spans education, health, consensual adult expression, pornography, exploitation, sex work, and content involving minors. Different jurisdictions draw boundaries in radically different places and impose different obligations: age verification, obscenity tests, distribution bans, anti-trafficking rules, and advertising restrictions.

AI systems also introduce a special sensitivity: the boundary between descriptive content and explicit content, between advice and arousal, between education and production. Even if a jurisdiction is permissive, a provider may treat explicit generation as high-risk because it is difficult to reliably keep minors out, prevent coercion, prevent non-consensual content, and prevent exploitation. This produces a familiar pattern: educational, clinical, or relational content may be allowed, while explicit content is refused; or the reverse in some markets where public health speech is more restricted.

Failure modes include: overblocking of legitimate sex education, inconsistent gating, and "false purity" behavior where the system refuses terms but cannot offer safe alternatives.

3.4 LGBTQ speech and identity: legality, “propaganda,” and criminalization risk

This domain fragments not because identity itself is complex, but because states and cultures differ sharply on what counts as permissible public speech about identity. In some jurisdictions, LGBTQ identity is treated as protected status, and anti-discrimination norms push systems toward supportive language. In others, public “promotion” is restricted, and in some cases identity-related organization is framed as forbidden or extremist. This is one of the clearest examples of “the map enters the interface”: a sentence that is normal affirmation in one place can be framed as illegal advocacy in another.

Providers typically treat this domain with a high “harm sensitivity” posture because it involves targeted harassment risk and because systems can easily be pulled into demeaning language or coercive therapy narratives. The fragmentation here is partly legal, partly moral, and partly platform-driven.

Failure modes include: route confusion (wrong jurisdiction overlay applied), “tone wars” where refusal or support is interpreted as ideology rather than policy, and chilling effects where users cannot access neutral informational content because the system cannot safely distinguish inquiry from advocacy under restrictive regimes.

3.5 Hate speech, group vilification, and extremist propaganda

This fractures because speech law and social norms differ radically. Some jurisdictions protect offensive speech but prohibit direct incitement; others criminalize group insult, denialism, or certain ideological symbols. Providers also impose global rules to reduce harassment and violence risk. In addition, “hate speech” is not always agreed upon: groups disagree on which language is violence-adjacent, which claims are defamation, and which criticisms are legitimate political speech.

This domain is also where “neutrality” becomes impossible. If a model repeats a slur, it is blamed. If it refuses, it is accused of censorship. If it provides a balanced summary of hateful ideology for analysis, it risks being used as propaganda. If it condemns, it risks moralizing and inflaming.

Failure modes include: asymmetry accusations (users perceive uneven enforcement), quotation traps (users ask the model to reproduce prohibited slogans), and the “analysis laundering” problem where propaganda is requested under the guise of scholarship.

3.6 Terrorism, insurgency, and “material support” boundaries

This category fragments because many jurisdictions treat certain kinds of information as facilitation. The same technical detail can be considered academic in one place and “instructional” in another. In addition, the classification of groups as terrorist or insurgent is

jurisdiction-dependent. One state's "freedom fighters" are another's "terrorists." This creates legal and moral collision.

The "material support" boundary is a key driver: guidance that helps a designated group recruit, fundraise, operationalize, or evade enforcement can be criminalized. Providers therefore tend to adopt strict safety floors on operational content. But users also have legitimate reasons to ask: journalists, researchers, historians, and citizens trying to understand conflict.

Failure modes include: overblocking legitimate analysis, underblocking when the model mistakes operational planning for history, and designation mismatch where the system's list of prohibited organizations differs from local law or user expectation.

3.7 Elections, persuasion, and targeted political messaging

This domain fragments because elections are the core legitimacy mechanism in many systems and a core stability concern in others. Regulations vary on political advertising, targeting, misinformation, foreign influence, campaign finance, and disclosure. AI adds a new risk: scalable persuasion and micro-targeting.

Providers often restrict explicit "write me a persuasive message for X demographic in Y district" requests, while allowing general civic information, candidate comparison, or policy explanation. The line between "education" and "persuasion" is thin. A user can request "neutral phrasing" and then deploy it as propaganda. Some regimes demand restrictions; others demand openness; still others prefer state-controlled messaging.

Failure modes include: selective enforcement perceptions, "both-sides" pressure, and the collapsing of legitimate activism into prohibited manipulation because the model cannot infer intent reliably.

3.8 Dissent, protest logistics, and criticism of state institutions

This category fractures because in many jurisdictions, protest is protected, while in others it is licensed, restricted, or suppressed. A model that provides practical protest logistics (how to organize, how to avoid detection, how to evade police tactics) can be framed as enabling disorder or subversion. Yet in open societies, "how to protest safely and legally" is legitimate civic guidance.

A jurisdiction-aware system must distinguish between: lawful assembly guidance, safety and de-escalation advice, and evasion/operational resistance. The distinction is often impossible at the prompt level. Users can mask intent, and the system cannot safely assume. Owners also avoid this domain because it can trigger government retaliation and reputational risks in multiple markets.

Failure modes include: chilling effects on legitimate civic speech, overbroad refusals that treat criticism as a threat, and inconsistent behavior that feels politically selective.

3.9 Surveillance, biometrics, and policing technologies

This domain fractures because privacy and security tradeoffs vary by regime. Some jurisdictions heavily restrict biometric identification and surveillance deployments; others actively pursue them as tools of governance. Users ask for everything from face recognition pipelines to phone tracking to OSINT. These requests can be framed as safety or as stalking, as lawful investigation or as oppression.

Providers often treat operational surveillance guidance as high-risk because it can enable abuse. Yet research and compliance users do exist: auditors, policymakers, privacy officers, journalists. The system must balance. It must also avoid becoming a "capability exporter" for regimes that use surveillance for repression.

Failure modes include: misclassification (blocking benign compliance questions), partial compliance (answering enough to enable misuse), and jurisdiction overlay conflict (local law allows one thing, provider policy forbids it).

3.10 Privacy, doxxing, and re-identification through aggregation

Privacy fragmentation is driven by differing legal regimes (data protection, defamation, public records) and differing cultural expectations. AI makes privacy harder because it can synthesize identity from many small signals: a "soft doxxing" pipeline. Users may ask for "who is this person," "find their address," "connect these accounts," "identify this voice," "pull everything public." Even if data is public somewhere, aggregation can be harmful.

Providers tend to enforce strict prohibitions on doxxing and personal data exploitation. But users still want legitimate uses: journalism, fraud prevention, verifying a business, confirming a public official's background. Jurisdiction overlays matter because some places allow public record aggregation, others restrict it, and defamation standards differ.

Failure modes include: overly permissive identification, overly restrictive refusal to discuss legitimate public figures, and inconsistent treatment of "public" information that feels arbitrary.

3.11 Cybersecurity, malware, and dual-use technical detail

This domain is dual-use by nature. The same knowledge can help defenders or attackers. Jurisdictions vary on cybercrime enforcement, but the bigger driver is provider policy: avoid enabling harm. Users often ask for: exploit code, phishing templates, credential stuffing tips, bypass methods, malware obfuscation, and "how to hack." Others ask for defensive guidance: patching, threat modeling, secure coding.

A good system needs a tiered response strategy: allow defensive principles, allow safe demonstration in toy settings, refuse actionable exploit instructions, and route to best practices. But fragmentation persists because different platforms and markets have different appetites for how much “red team” content is allowed.

Failure modes include: leakage of actionable steps through “educational framing,” and overblocking where the system refuses even basic defense questions.

3.12 Medical advice, reproductive health, and controlled medications

Medical content fragments because of licensing boundaries, malpractice exposure, and politicized health topics. Users ask for symptom interpretation, dosage advice, drug interactions, how to obtain medications, and self-treatment plans. The system must distinguish education from medical practice, and must avoid replacing clinicians. In addition, reproductive health is highly politicized in many jurisdictions, which can lead to differing constraints on what can be said.

Controlled medications add another layer: substance schedules, prescription requirements, and procurement restrictions. Even when someone requests legitimate information, the system may refuse because it cannot verify intent or identity.

Failure modes include: false reassurance, inappropriate specificity, refusal of legitimate harm-reduction or “when to seek care” guidance, and jurisdictional confusion on what is legal or available.

3.13 Legal advice: firearms/drugs transport, immigration, criminal exposure

Legal topics fragment because law is not portable. People ask AI for compliance: what to do, what is risky, how to respond to police, how to cross borders with regulated items, how to handle visas, how to plead, how to structure a defense. The system sits near the unauthorized practice of law boundary and near high-stakes personal consequence.

Providers typically adopt broad disclaimers and conservative behavior: general information only, encourage professional counsel, avoid strategic advice for wrongdoing. But users want specificity. The gap creates frustration, and it creates fragmentation: some jurisdictions and products allow more detailed legal summarization, others restrict it.

Failure modes include: confident wrong answers that create legal harm, vague answers that are useless, and accidental facilitation of evasion framed as “rights information.”

3.14 Finance: crypto, gambling, sanctions, and evasion-adjacent requests

Finance fragments because regulatory regimes diverge: securities law, consumer protection, gambling legality, AML/KYC requirements, and sanctions enforcement. Crypto adds volatility: some places embrace it, some restrict it, some ban it, and enforcement varies.

Users ask for investment recommendations, tax strategies, how to move money across borders, how to avoid reporting, how to use mixers, how to evade sanctions, and how to exploit regulatory gaps. The provider must avoid enabling evasion or fraud and avoid giving personalized financial advice. Yet legitimate questions exist: understanding risk, compliance, tax basics, and general market education.

Failure modes include: "advice laundering" where evasion is asked as "privacy," and inconsistent handling across markets and product tiers.

3.15 Religion and blasphemy constraints in specific jurisdictions

Religion fragments because some jurisdictions criminalize blasphemy, apostasy, or religious insult, while others protect robust criticism of religion as speech. Even where law is permissive, social enforcement can be severe. AI systems become flashpoints because they can generate satire, critique, and comparative analysis at scale.

Users also ask the model to adjudicate doctrine, to produce polemics, to attack other faiths, or to "prove" superiority. Providers often restrict harassment and hate content, but religious critique occupies a complicated boundary: what is critique, what is vilification, what is incitement?

Failure modes include: chilling effects on legitimate scholarship, selective suppression perceived as favoritism, and model escalation where it produces inflammatory content because it is following an adversarial instruction.

Across all these categories, one pattern repeats: high fragmentation is not primarily about "opinions." It is about whether the system's outputs can be interpreted as enabling illegal action, destabilizing regimes, facilitating harassment, or causing health and safety harm. That is why these topics fracture first. They are where words become instruments, and where jurisdictions and owners insist on controlling the instrument.

The next sections will build on this taxonomy by describing the drivers of divergence in a more mechanistic way, then detailing the failure modes, then proposing design patterns that make the system's constraints legible rather than arbitrary.

4. Drivers of Divergence: Why the Same Prompt Breaks Differently

Users often experience fragmentation as a mystery: “Why does the model answer this for my friend but not for me?” “Why did it answer last month but refuse today?” “Why does the same question become ‘unsafe’ when I cross a border?” These experiences feel personal because the interface is personal. Yet the driver is usually impersonal: the prompt has entered a different constraint environment.

This section explains why the same prompt breaks differently. “Break” here includes refusals, hedging, downgraded detail, altered tone, or a shift from direct answer to generic safety talk. The core idea is that “what the user asked” is not the only determinant. The system also classifies intent, estimates risk, selects a policy profile, and applies jurisdictional overlays. Divergence arises when any of those elements changes across place, platform, or time.

The drivers below are the main reasons. They are phrased as variances because fragmentation is, in effect, a function of variance: legal variance, enforcement variance, institutional variance, and cultural variance.

4.1 Criminal law variance: definitions, thresholds, and intent standards

Criminal law variance is the most concrete driver because it produces the clearest “cliff edge” behavior: allowed here, illegal there. But the variance is rarely just a binary. It typically appears in three forms: definitions, thresholds, and intent standards.

Definitions differ. What counts as a “weapon,” “concealed,” “distribution,” “controlled substance,” “obscene material,” “hacking,” or “threat” can change across jurisdictions. Even when two places use the same term, their case law and statutory definitions diverge. This matters because a prompt that seems neutral may be interpreted as advice about a defined offense. A model that cannot reliably map user language onto local definitions will default to conservative behavior.

Thresholds differ. Many criminal categories change severity based on quantity, proximity, age, location, or prior offenses. Small possession may be treated as a misdemeanor, while quantities above a threshold trigger presumptions of distribution. Carrying a regulated item may be lawful generally but illegal within a school zone, near a government building, or while intoxicated. Cyber offenses often have thresholds around unauthorized access, intent to impair, or value of losses. The same user request can be a harmless “overview” in one place and a blueprint for crossing a threshold in another.

Intent standards differ. Criminality often turns on mental state: knowingly, recklessly, with intent to distribute, with intent to evade law enforcement, with intent to harass, with intent to defraud. The model cannot read minds. It infers intent from the prompt, user history, and

pattern matching. But intent inference is brittle: users may not state their real purpose, and legitimate purposes can resemble illegitimate ones. Jurisdictions also vary in how they infer intent from circumstances. A model that tries to be "helpful" under uncertainty can be accused of enabling crime. So owners bias toward refusal or high-level answers when intent inference is weak.

The combined effect is a systematic divergence. The prompt is the same, but the underlying risk classification differs because the local criminal law landscape changes what the prompt could plausibly enable. When in doubt, the system behaves as if the harshest interpretation is relevant, because the downside of being wrong is asymmetric.

4.2 Speech law variance: illegal content categories and duty-of-care regimes

Speech law variance is often misread as "censorship bias," but the mechanism is more specific. Different jurisdictions define different categories of illegal or regulated speech, and some impose affirmative duties on platforms and intermediaries. This creates two types of divergence: content category divergence and procedural divergence.

Content category divergence means that certain utterances are illegal in some places and protected in others. These categories can include group insult, denialism, threats, extremism, misinformation, defamation, blasphemy, obscenity, promotion of prohibited behavior, or publication of private information. The exact category list varies widely and is politically charged. A model that speaks uniformly will violate law somewhere. So systems either localize their behavior or withdraw.

Procedural divergence is about duties. Some regimes demand prompt takedown, proactive moderation, age verification, traceability, retention of user data, or risk assessment documentation. These obligations shape AI behavior even when the content itself is not illegal. A provider may implement conservative refusals to reduce compliance burden, because every borderline output increases the probability of a regulator dispute. Thus, the divergence can show up not as "you may not say X" but as "we will not risk generating content that might be categorized as X."

Duty-of-care regimes intensify the strictest-layer effect. When platforms face penalties for failing to prevent harm, they tend to overcomply. Overcompliance is not ideology; it is a predictable response to liability and enforcement uncertainty. The user experiences it as a speech shift. The underlying driver is the platform's legal environment.

4.3 Security variance: "state stability" doctrines and forbidden narratives

Security variance is the driver that most directly converts "talk" into "threat." In many jurisdictions, certain narratives are not treated as debate; they are treated as destabilization.

This includes topics like protests, separatism, leader legitimacy, war narratives, historical trauma, ethnic tensions, and allegations of state wrongdoing.

“State stability doctrine” is not always written as one statute. It can be a patchwork of laws on extremism, public order, rumor spreading, sedition, national security, insulting state symbols, or foreign agent activities. But even more important is enforcement doctrine: what the state actually prioritizes.

For AI deployments, security variance yields two effects. First, localization demands: the state wants outputs aligned with its stability doctrine. Second, risk-based withdrawal: the provider may avoid operating in that environment if compliance would compromise global integrity or create reputational blowback elsewhere.

For users, the most confusing manifestation is narrative asymmetry. The model may discuss sensitive issues in one region and refuse or sanitize in another. Users interpret this as political capture. Often, it is. But the mechanism is not hidden. It is rooted in differing state requirements and the owner’s tolerance for adopting them.

Security variance also includes “foreign influence” sensitivity. Some states treat certain content as evidence of hostile foreign manipulation. That creates a feedback loop: the more a system is seen as “foreign,” the more it is pressured to demonstrate controllability and alignment with local narratives.

4.4 Privacy variance: data localization, sensitive data, and identification rules

Privacy variance is a driver because “privacy” is not one universal concept. It ranges from strong individual rights (limiting collection, processing, and retention) to sovereignty-driven data control (keeping data within national borders), to security-prioritized systems where identification and traceability are treated as essential.

Data localization is a structural divergence. If a jurisdiction requires that user data remain within its borders, a cloud-deployed foreign model may be incompatible unless the provider offers local hosting. This can force different model variants, different logging, different retention, and different tool access. It is not just “policy.” It is architecture.

Sensitive data definitions differ. What counts as sensitive (health, biometrics, political opinions, religious beliefs, sexual orientation, children’s data) differs across regimes and is treated with different protections. A model that can infer or repeat sensitive attributes creates special risk. So systems may refuse or sanitize in ways that are highly jurisdiction-dependent.

Identification rules differ. Some jurisdictions treat the ability to identify individuals as a privacy harm; others treat identification as a security necessity. This affects whether a model can assist with facial recognition pipelines, OSINT aggregation, or linking accounts across platforms. It also

affects the system's own user management: anonymous use versus verified accounts, logging depth, and response traceability.

The net effect is that the same prompt can become a different category of risk based on privacy regime. "Help me identify this person" can be framed as investigative journalism in one context, stalking in another, and as a prohibited identification request under a strong privacy regime. The system's behavior diverges accordingly.

4.5 Civil liability variance: negligence, defamation, and product safety exposure

Even when criminal law does not prohibit an output, civil liability can. Owners are often more sensitive to civil exposure than to criminal exposure because civil risk is common, frequent, and expensive. The civil liability driver is especially important in open societies where speech is protected but lawsuits are plentiful.

Negligence is the first pathway. If a model gives instructions that lead to foreseeable harm, plaintiffs may claim the provider failed to exercise reasonable care. This is the "duty and breach" logic. Providers respond by refusing operational detail in high-risk categories and by offering general information and warnings instead.

Defamation is the second pathway. Models can generate false claims about individuals, businesses, or groups. Defamation standards differ across jurisdictions, including burdens of proof, protections for opinion, and special rules for public figures. The safer path is to avoid making factual allegations about identifiable persons or to rely on reputable sources. That produces divergence: the same prompt about an individual may be answered cautiously in one jurisdiction and more freely in another, depending on defamation risk.

Product safety and consumer protection are the third pathway. If the AI is treated as a consumer product, regulators may view harmful outputs as product defects or deceptive practices. This can force providers into conservative behavior, disclaimers, and structured "do not rely" messaging. Again, the divergence is not about truth. It is about exposure.

Civil liability variance produces a globalizing effect: owners may adopt the most conservative posture globally to minimize litigation risk anywhere. This is one of the main mechanisms through which local variance turns into global blandness.

4.6 Cultural variance: social enforcement, not just legal enforcement

Cultural variance is the driver that makes fragmentation persist even when laws are similar. Culture includes norms, taboos, reputational penalties, community enforcement, employer consequences, family sanctions, and informal power dynamics. It often governs what users dare to ask, what they interpret as offensive, and what they demand the system refuse.

Social enforcement can be more powerful than legal enforcement. A topic may be legal but taboo, leading to strong backlash if a platform is seen as "promoting" it. Or a topic may be illegal but socially accepted, leading users to demand help despite the law. The model becomes a point of conflict not because the content is inherently controversial, but because the community's identity is tied to policing that content.

Cultural variance also changes how refusal is read. A refusal can be read as moral condemnation in one culture and as normal safety practice in another. A neutral factual tone can be read as cold disrespect in one place and as professional restraint in another. Even the choice of words—clinical versus relational, direct versus indirect—can trigger perceptions of bias.

This is why localization is not only about restricting content. It is also about adjusting style, disclaimers, and the framing of permissible alternatives. A system that is legally compliant can still fail culturally. Owners often respond by simplifying outputs, avoiding nuance, or withdrawing from certain conversations. That produces fragmentation that users experience as "the AI has a worldview," when the underlying driver is cultural enforcement pressure.

Taken together, these drivers explain why "the same prompt" is an illusion. A prompt is not evaluated in isolation. It is evaluated inside a living constraint environment: criminal law, speech law, security doctrine, privacy regime, civil liability, and culture. Change any of these, and the output may change—even if the weights are identical.

This framing sets up the next section: if divergence drivers are predictable, then the failure modes are predictable as well. We can name them, measure them, and design against them rather than treating them as mysterious bias or random inconsistency.

5. Failure Modes: What Fragmentation Looks Like to Users

Fragmentation is not experienced as a legal doctrine or a governance architecture. It is experienced as a conversational rupture. Users do not see the policy router or the jurisdiction pack. They see a sentence that was answered yesterday and refused today. They see a request answered for someone else but not for them. They see a model that sounds helpful in one domain and suddenly becomes evasive, preachy, or bureaucratic in another. These surface experiences matter because they define trust. If fragmentation is unmanaged, it trains users to interpret the system as arbitrary power.

This section names the main failure modes at the user interface level. The goal is not to blame any one actor. The goal is to identify patterns that are predictable once a system is governed by

multiple sovereignties. Each failure mode below is a stable class of user-visible dysfunction. If you can name it, you can measure it. If you can measure it, you can engineer mitigations.

5.1 Overblocking: safe information refused because routing is too blunt

Overblocking occurs when the system refuses safe, lawful, or educational content because the risk classifier is too coarse. The system's routing logic sees a keyword or topic region associated with harm and triggers a blanket refusal. Users experience this as senseless prudishness or as an ideological veto. In reality, it is often a mechanical artifact: blunt routing deployed to avoid rare but catastrophic misuse.

Overblocking has several common causes. One is keyword-based gating. If the policy router uses simple triggers, the presence of a term like "gun," "cannabis," "abortion," "explosive," "hack," or "porn" can activate a high-risk policy profile even when the user's request is safety-oriented or legal. Another cause is low-confidence intent inference. If the system cannot confidently distinguish "compliance question" from "evasion question," it may refuse both. A third cause is policy compression: the provider chooses one globally safe rule instead of maintaining nuanced jurisdiction-aware distinctions.

The harm of overblocking is not only user frustration. It can push users toward riskier sources, reduce harm-reduction education, and create a culture of adversarial prompting. When users feel blocked, they try to outsmart the system. This increases jailbreak pressure and paradoxically increases risk.

Overblocking is also reputationally expensive because it looks like selective censorship. When a system refuses basic, non-operational questions, users infer bias. If the provider cannot explain the layer causing the refusal, the refusal becomes a Rorschach test for political interpretation.

5.2 Underblocking: harmful operational detail slips through in "neutral" framing

Underblocking is the mirror image: the system provides actionable harmful guidance because the request was framed in a way that bypassed triggers. The user learns that the model does not block "bad content"; it blocks certain patterns of phrasing. This is a deeper failure than overblocking because it violates the safety floor while simultaneously encouraging adversarial prompting.

Underblocking often happens through "neutral framing." A user requests operational violence or illegal procurement, but wraps it in academic language: "for research," "for a story," "for historical interest," "for a school project," "for defense." Because the model is trained to be helpful and to follow instructions, it may comply with the literal request while missing the operational threat. Underblocking also arises when the system provides "partial steps" that, when combined with external knowledge, become fully actionable. It can also arise when the

model supplies high-level explanation plus enough implementation detail to cross the facilitation boundary.

Jurisdiction fragmentation worsens underblocking because different overlays may weaken or strengthen guardrails. A model might be tuned to refuse certain instructions globally, but an alternate product tier or a localization pack might unintentionally loosen constraints. Underblocking is therefore not only a model issue; it is a systems integration issue.

To users, underblocking looks like hypocrisy. The model refuses innocuous questions, then provides dangerous detail when asked "the right way." This breaks trust in the idea that refusals are principled. Users conclude that policy is theater, and once they believe that, they stop cooperating with safety constraints.

5.3 Self-sealing refusals: non-falsifiable policy language that escalates distrust

A refusal can be correct and still produce backlash if it is communicated badly. Self-sealing refusals are refusals that cannot be contested because they provide no testable reason. The refusal becomes a closed loop: "I can't help with that" plus an implied moral judgment, with no indication of which rule was invoked, what aspect of the request triggered it, and what alternative information is safely available.

Self-sealing refusals escalate distrust for two reasons. First, they feel arbitrary. If the user cannot see the boundary, they cannot adapt their request. Second, they resemble political gatekeeping. When a refusal is framed as "harmful" without specificity, the user may infer that the model is punishing disfavored viewpoints rather than preventing concrete misuse.

Self-sealing refusals also generate conspiracy dynamics. When the system refuses without reasons, users fill the vacuum with narrative. They attribute the refusal to shadowy influence, hidden policy agendas, or ideological capture. The refusal becomes "evidence" of control precisely because it refuses to explain itself.

The antidote is not to reveal sensitive policy details that enable bypass. The antidote is to offer a bounded explanation: identify the relevant risk category, name the layer (safety floor versus jurisdiction overlay versus product policy), and provide safe alternatives (high-level overview, legal resources, safety guidance, or "consult a licensed professional"). The goal is contestable clarity without leakage.

5.4 Jurisdictional misrouting: wrong legal regime applied to the user

Jurisdictional misrouting occurs when the system applies the wrong "map overlay" to a user's request. This can happen because the system inferred the user's location incorrectly, because the user is traveling, because the account region differs from the physical location, because a

VPN or proxy changed signals, or because the system applied a default region due to missing data.

Misrouting produces two types of user harm. The first is false permission: the model suggests that something is lawful or safe based on the wrong jurisdiction. Users who follow this guidance can face legal consequences. The second is false prohibition: the model refuses content because it thinks the user is in a restrictive jurisdiction when they are not. Users experience this as unjust restriction and often escalate into adversarial behavior.

Misrouting also creates reliability collapse. Users stop trusting the system even on neutral topics because they see it as incapable of grounding its advice in the correct context. In legal and health domains, misrouting is particularly damaging. A small jurisdictional error can flip a user's understanding of risk.

A robust system needs a clear policy for jurisdiction uncertainty. It should default to non-actionable, high-level information when location is uncertain, and it should encourage users to verify local law through official resources. Importantly, it should avoid "confidently local" answers unless it is confident about location and the relevant statute environment.

5.5 Asymmetric enforcement: perceived favoritism caused by unequal constraint sets

Asymmetric enforcement is what users call "bias," but it can arise even when the underlying cause is structural. It occurs when the system seems to apply stricter rules to some topics, ideologies, or groups than to others. The user may conclude that the model is protecting certain narratives or punishing certain viewpoints.

There are several sources of asymmetry. One is that different jurisdictions and platforms genuinely restrict different content categories. If a system is deployed globally, it may become stricter on topics that trigger legal risk in one major market, even if users elsewhere view those topics as ordinary speech. Another source is uneven detectability. Some harmful content is easier for classifiers to detect than other harmful content, leading to uneven enforcement simply because one category triggers more reliably. A third source is policy layering: the safety floor might be symmetric in principle, but the owner policy adds exceptions, or the jurisdiction pack adds restrictions that apply more often to certain content.

Users rarely see these mechanisms. They see outcomes. And outcomes are compared socially. If people can share screenshots, they can build a case that "the model allows X but not Y." Whether the case is accurate or cherry-picked, the perception can become durable.

Asymmetric enforcement is especially dangerous because it corrodes legitimacy. Users will accept constraints if they believe constraints are applied consistently. They will resist constraints

if they believe the system is playing favorites. This is why transparency about layers matters: not to excuse inconsistency, but to make it diagnosable and correctable.

5.6 The whiplash effect: crossing a border changes the model's persona

Whiplash is the most visceral fragmentation experience. A user travels or changes networks, and the model's behavior shifts abruptly: it becomes more cautious, more moralistic, more evasive, or more restrictive. Sometimes it becomes the opposite: more permissive, more blunt, more opinionated. The same user feels as if they are talking to a different system.

Whiplash has several causes. One is literal geofencing: features and policies differ by region. Another is jurisdiction packs: the local overlay changes how the model is allowed to respond. A third is product tier differences that correlate with region: enterprise deployments may have different settings than consumer deployments, and those deployments are unevenly distributed geographically. A fourth is language and culture: the model may adopt different tone conventions in different languages, which users interpret as a personality shift.

Whiplash damages trust because it reveals the system as governed. Users may have been willing to imagine the model as an objective tool. Whiplash forces them to see it as an instrument subject to power. If this revelation happens without explanation, users interpret it as manipulation.

The design response is not to pretend whiplash can be eliminated. It is to make whiplash legible. The interface should disclose when a jurisdiction or policy profile changes. It should indicate that "this environment imposes additional constraints," and it should allow users to understand whether the change is legal, safety-based, or product-based. Without such disclosure, users will build their own explanations, and those explanations will often be adversarial.

These failure modes are not fringe. They are the predictable surface artifacts of a multi-sovereign system living under asymmetric downside. Overblocking and underblocking are the two poles of safety error. Self-sealing refusals are the communication failure that turns safety into distrust. Misrouting is the grounding failure that turns localization into danger. Asymmetric enforcement is the fairness failure that turns policy into politics. Whiplash is the experiential failure that turns the model into a map-dependent persona.

The next step is to move from naming to measurement and mitigation: how to quantify fragmentation, how to audit it across jurisdictions, and how to build "jurisdiction-aware" systems that preserve user dignity and contestability while still respecting safety floors and legal constraints.

6. Measuring Fragmentation: Metrics that Replace Vibes with Evidence

Fragmentation becomes politically poisonous when it is argued as a matter of vibe: “the model feels biased,” “it’s censoring,” “it’s captured,” “it’s reckless,” “it’s inconsistent.” Vibes are not nothing; they are user experience. But if fragmentation is treated only as vibe, then the loudest interpreters win and the system becomes ungovernable. What is needed is a measurement layer that turns fragmentation into something you can audit, compare, regress, and improve.

The purpose of measurement is not to prove neutrality. Neutrality is not a measurable target in a plural world. The purpose is to make divergence legible: where it occurs, why it occurs, whether it is justified, and how it changes over time. This section proposes a minimal metric suite that can be implemented without requiring the provider to expose sensitive internal policy details. The metrics are designed to be externally auditable in principle, at least in aggregate, and to support meaningful comparisons across regions, products, and model versions.

A useful mindset is that an AI deployment has two outputs: the conversational answer, and the “governance trace” that explains how the answer was assembled. Many fragmentation failures arise because the trace is missing. Metrics make the trace visible, even if only statistically.

6.1 Topic volatility index: how often rules differ across jurisdictions

The topic volatility index measures how unstable a topic is across jurisdictions and policy profiles. It answers a basic question: if the same intent is expressed in different places, how often do the system’s permissions diverge?

Define a set of topic families (as in Section 3) and a set of standardized prompt intents for each family. For each jurisdiction and product tier, record whether the response is: allowed, allowed with constraints, redirected (high-level only), or refused. Volatility is then a function of dispersion: how many distinct policy outcomes occur across jurisdictions for the same topic-intent pair.

A simple formulation is:

- For each topic-intent pair, compute the distribution of outcome classes across jurisdictions.
- Convert that distribution into an entropy-like score (higher entropy = more divergence).
- Aggregate across intents to get a topic-level volatility score.

High volatility topics will show wide dispersion: weapons operational detail, drug procurement, sexual content generation, protest logistics. Low volatility topics will show consistent outcomes: neutral history, math, general writing help.

The value of this index is that it predicts user pain. High volatility topics are where users most often experience “the map entered the interface.” If a provider knows which topics are volatile, it can prioritize disclosure, careful routing, and safe alternative pathways. If a regulator or auditor sees volatility, they can focus oversight on those fracture planes rather than arguing about everything at once.

Volatility also separates “political controversy” from “policy complexity.” A topic can be culturally controversial but low-volatility if the platform takes a uniform stance; or culturally calm but high-volatility if legal frameworks diverge sharply.

6.2 Refusal gradient mapping: where refusals cluster and why

Refusal gradient mapping is the spatial and contextual visualization of refusals. Instead of treating refusal as a binary event, it treats it as a gradient across three dimensions: topic, intent, and jurisdiction.

The first move is to define refusal types. Not all refusals are the same. At minimum, distinguish:

- Hard refusal (no content beyond generic refusal)
- Soft refusal (offers high-level info but no operational detail)
- Redirect refusal (points to official resources, professional help, or safe alternatives)
- Conditional refusal (would answer if constraints were met, e.g., age-gated or licensed context)

Then map refusal frequency and refusal type across jurisdictions and topics. The result is a “refusal landscape” that shows clusters: where the system becomes rigid, where it becomes permissive, and where it becomes inconsistent.

The “why” requires attribution. A refusal without attribution is politically interpretable. A refusal with attribution is auditable. Even if the system cannot reveal internal policy text, it can tag each refusal with a coarse reason code:

- Safety floor
- Jurisdiction pack
- Owner policy
- Uncertain intent
- Uncertain jurisdiction
- Tool limitation (cannot verify)

- Sensitive personal data risk

Now you can build gradient maps that show not just where refusals happen, but which layer is driving them. That allows targeted fixes. If refusals cluster due to uncertain jurisdiction, improve location confidence gating and disclosure. If refusals cluster due to vague safety floor rules, refine the “safe alternative” templates. If refusals cluster due to owner policy, accept that as a business choice but disclose it cleanly so users do not misread it as legal requirement.

Refusal gradient mapping also reveals overblocking and underblocking risks. Overblocking appears as broad refusal regions for benign intents. Underblocking appears as holes in the refusal surface for clearly harmful intents. The mapping makes those holes visible.

6.3 Parity tests: symmetrical handling of comparable claims about different groups

Parity tests address the perceived fairness problem: “the model enforces rules unevenly.” If a system treats comparable claims differently depending on the group involved, users will attribute ideology. Some differences are justified by local law or by differential harassment risk patterns, but many differences are accidents of training priors or classifier sensitivity.

Parity testing requires controlled prompt design. You create prompt templates that are identical except for the target group or entity. For example:

- Comparable praise requests (“write a flattering description of group X”)
- Comparable criticism requests (“write a harsh critique of group X”)
- Comparable stereotyping prompts (“what are typical traits of group X?”)
- Comparable political framing prompts (“why is group X dangerous/beneficial?”)

Then you measure outcomes: refusal rates, detail levels, tone markers, and whether the model generalizes from individuals to groups. The tests should be run across jurisdictions and across policy profiles because parity can change under localization packs.

Parity metrics should include:

- Outcome parity (allowed/refused/redirected distribution)
- Tone parity (sentiment and moralizing language)
- Generalization parity (how often the model makes sweeping claims)
- Evidence parity (how often the model requests sources or caveats)

The point is not to force identical behavior everywhere. The point is to detect unintended asymmetries that look like favoritism. If a provider wants legitimacy, it must be able to say: “we

run parity tests; we can show where differences are policy-driven versus accidental; we publish aggregate results."

Parity tests also serve as a safety mechanism. Many harmful narratives are "copy-paste adaptable." If the model blocks one group-targeted hate prompt but allows the same prompt with a different target, the system becomes a hate amplifier through loopholes. Parity testing closes that loophole.

6.4 Compliance drift tracking: when policies change silently over time

One of the deepest trust failures in AI is silent change. Users return to the same question and get a different answer, not because the world changed, but because policies or model versions changed. If those changes are not disclosed, users infer manipulation. They also lose the ability to reproduce outputs, which undermines scientific and legal uses.

Compliance drift tracking measures how the system's behavior changes over time across the same standardized test set. The measurement must track both:

- Model drift (weights, tuning, and capability changes)
- Policy drift (routing rules, safety rules, jurisdiction packs, owner policy)

A robust system should publish version identifiers and policy profile identifiers that are included in the governance trace. Drift tracking then becomes straightforward: run the same evaluation suite across versions and record differences.

Key drift metrics include:

- Refusal drift (changes in refusal rates by topic-intent)
- Detail drift (changes in allowed specificity)
- Attribution drift (changes in reason codes)
- Jurisdiction drift (changes in which packs apply where)
- Explanation drift (changes in refusal phrasing and alternatives)

Drift can be legitimate. Laws change. Incidents occur. New safety vulnerabilities are discovered. But legitimacy requires disclosure. Drift tracking makes it possible to say: "On this date, policy changed; here is the class of behaviors impacted; here is why." Without drift tracking, "policy updates" become indistinguishable from ideological steering.

Drift tracking also supports governance accountability. If a provider claims to uphold a safety floor but drift data shows increasing underblocking in a dangerous category, that is actionable. If

a provider claims to maintain parity but drift data shows growing asymmetry, that is also actionable.

6.5 User trust metrics: contestability, clarity, and appeal success rates

Ultimately, fragmentation is experienced as a trust event. Measurement must therefore include user-centered trust metrics that reflect whether users feel respected, informed, and able to contest decisions.

Contestability measures whether a user can understand and challenge an outcome. A refusal that includes a reason code and safe alternatives is more contestable than a blanket "no." A system that supports an appeal pathway is more contestable than one that does not.

Clarity measures whether the system's explanations are understandable and actionable. Clarity is not verbosity. It is whether the user can tell: what triggered the constraint, which layer is acting, and what is permitted instead. Clarity can be measured through user surveys, task completion rates, and "prompt repair" success (how often users successfully rephrase a request into an allowed form without adversarial behavior).

Appeal success rate measures whether users can obtain corrected outcomes when the system overblocked or misrouted. An appeals system can be as formal as a support ticket or as lightweight as an in-chat "I think this is allowed; here's my context" flow. Appeal success can be measured as:

- Percentage of refusals reversed after context clarification
- Time-to-resolution
- Rate of repeat refusals on the same issue
- User drop-off after refusal

High appeal success suggests the system is conservative but responsive. Low appeal success suggests either the policy boundary is absolute or the system is too blunt. In either case, the user experience differs. Importantly, appeals also provide training data for improving routing and reducing overblocking.

Trust metrics should also include whiplash detection: how often the system's tone and refusal style changes across sessions, regions, or devices. If whiplash correlates with location changes, the interface should disclose that explicitly. If whiplash correlates with silent policy updates, the provider should surface version changes.

These measurement tools replace interpretive conflict with diagnosable behavior. The topic volatility index tells you where fragmentation is structurally expected. Refusal gradient mapping shows where and why refusals cluster. Parity tests detect unfair asymmetries and loopholes. Compliance drift tracking preserves reproducibility and accountability. User trust metrics ensure that governance does not become humiliation.

Once this metric suite exists, the system can evolve under evidence rather than outrage. Without it, the model becomes a political object by default, because every refusal and every inconsistency becomes raw material for competing narratives. The next sections will build on this by proposing design patterns and governance mechanisms that make these metrics actionable: modular policy layers, disclosure interfaces, audit logs, and contestable decision pathways.

7. Design Patterns for Jurisdiction-Aware AI

If fragmentation is inevitable, the engineering question becomes: how do you build systems that are safe, lawful, and useful without turning the user experience into arbitrary power? Jurisdiction-aware AI is not primarily a better classifier or a stricter filter. It is a design approach: modular constraints, explicit provenance, honest disclosures, and an auditable governance trace.

The guiding principle is separation of concerns. Today's systems often blur four things into one opaque behavior: capability (what the model could say), permission (what it may say), packaging (how it presents what it says), and provenance (why it behaved that way). When these are fused, users cannot tell whether a refusal is about safety, law, business policy, or uncertainty. That confusion is what makes the system politically fragile.

The patterns below are meant to produce three outcomes simultaneously: (1) maintain a non-negotiable harm floor, (2) satisfy genuine legal differences across jurisdictions, and (3) preserve contestable, dignified user interaction by making constraints legible.

7.1 The global safety floor: invariants that should not vary

A jurisdiction-aware system still needs a moral and safety spine. Without invariants, the system becomes a shapeshifter that can be bent to any regime or market. The safety floor is the set of constraints that travel with the model regardless of jurisdiction, because they represent the provider's minimum ethical and risk commitments.

The safety floor should be small, explicit, and defensible. The error most providers make is inflating the floor into a giant catch-all that becomes indistinguishable from ideology. A better

approach is to restrict invariants to clear, high-severity harm enablement categories where "helpfulness" plausibly becomes facilitation.

A practical safety floor usually includes: operational violence assistance, instructions for weaponization, guidance enabling fraud and theft, instructions for malware and unauthorized access, doxxing and personal data exploitation, sexual exploitation content (especially anything involving minors), and self-harm facilitation. These are not "controversial opinions." They are high-risk enablement pathways. The invariance claim is: we do not help users cause severe harm, regardless of local permissiveness or local cultural norms.

Two properties make a safety floor legitimate. First, it must be coupled to measurable harm models and red-team evidence, not just intuition. Second, it must be phrased in operational terms: "we do not provide instructions that increase capability to commit X," rather than "we do not support ideology Y."

The safety floor must also have a disciplined alternative mode: when refusing operational harm, the system should offer safe substitutes such as de-escalation, legal safety information, crisis resources, or general education that does not enable action. Otherwise, refusals become humiliation rather than protection.

7.2 Jurisdiction packs: modular legal overlays with explicit provenance

Jurisdiction packs are the core mechanism for managing legal variance without rewriting the model every time. The pack is a modular overlay that implements jurisdiction-specific constraints, obligations, and required disclosures. It is not merely a list of banned words. It is a structured policy module.

A robust jurisdiction pack includes at least five components:

First, content constraints: specific categories of content that are illegal or regulated in that jurisdiction (speech categories, controlled substances, weapons restrictions, privacy protections, etc.). These constraints should be expressed as policy rules that can be tested and updated.

Second, procedural obligations: requirements about age gating, logging, takedown responsiveness, retention, data localization, and reporting. These obligations often drive architecture choices, not just content filtering.

Third, escalation pathways: what the system must do when encountering certain categories (refuse, provide high-level info, refer to official resources, or require professional involvement).

Fourth, provenance metadata: the pack must carry identifiers such as jurisdiction code, version number, effective date, and a human-readable summary of what it changes. This enables drift tracking and reproducibility.

Fifth, conflict resolution rules: how the pack interacts with the global safety floor and with owner policy. For example, the pack cannot authorize what the safety floor forbids, but it can further restrict what the safety floor permits.

Explicit provenance matters because users are not just asking for answers; they are asking for predictable behavior. If the system's behavior changes because a jurisdiction pack changed, the system must be able to say so. If the system is applying a jurisdiction pack due to inferred location, it must be able to disclose that inference and its confidence level.

A crucial design choice is to treat jurisdiction selection as a first-class, inspectable state, not a hidden assumption. Hidden jurisdiction is the seed of misrouting and distrust.

7.3 Use-case packs: enterprise, education, clinical, newsroom, research modes

Jurisdiction is not the only axis of legitimate divergence. The same model can be deployed in contexts with different duty-of-care expectations and verification capabilities. A hospital deployment can verify user identity and enforce clinical governance. A newsroom deployment can enforce editorial review. An educational deployment can enforce age gating and curriculum constraints. A consumer chat app cannot reliably do any of these.

Use-case packs are therefore necessary. They are modular overlays keyed to the deployment environment, not to geography. A use-case pack defines permissible depth, tooling access, and escalation rules appropriate to the context.

Typical packs include:

Enterprise pack: assumes trained users and compliance controls. May allow deeper technical discussion in safe domains, with strict logging and access controls.

Education pack: prioritizes age-appropriate content, pedagogical explanation, and safeguarding. It may restrict explicit content and certain persuasive content, while emphasizing critical thinking.

Clinical pack: restricts personalized diagnosis and treatment planning; emphasizes triage guidance, patient education, and clinician-in-the-loop workflows. It demands high reliability and conservative uncertainty handling.

Newsroom pack: emphasizes sourcing, careful attribution, and avoidance of defamation; encourages verification and offers structured summaries rather than definitive allegations.

Research pack: allows more technical depth under strong guardrails, often with explicit red-team constraints and logging, and with clear prohibitions on operational misuse.

These packs solve a key legitimacy problem: users stop arguing about why the consumer app cannot do what the enterprise deployment can. The answer becomes structural, not moral: different use-case packs exist because different contexts can carry different risk.

Use-case packs also allow a system to remain useful without expanding the global safety floor. Instead of saying “no” everywhere, the system can say: “this level of detail requires a verified context.” That preserves safety while preventing the lowest-common-denominator collapse.

7.4 Policy routers: intent detection, location detection, and conservative fallbacks

The policy router is the runtime decision engine that selects which packs apply and how to respond. It is where fragmentation becomes either orderly or chaotic.

A policy router does three hard things under uncertainty:

Intent detection: determine what the user is trying to do, not merely what words they used. The router must distinguish education from facilitation, compliance from evasion, journalism from harassment, critique from vilification. This is inherently error-prone. Therefore, good routers treat intent detection as probabilistic. They do not pretend to know. They use confidence thresholds.

Location detection: determine which jurisdiction pack applies. This can be inferred from account region, IP geolocation, device settings, or explicit user declaration. Each signal can fail. Therefore, good routers compute a location confidence and have a policy for low-confidence states.

Conservative fallbacks: decide what to do when confidence is low. A router that guesses wrong can cause legal harm. The safe default is to provide non-actionable high-level information plus verification pointers, and to avoid operational guidance. The router can also prompt for clarification in low-stakes ways (“Which country or state are you asking about?”) but it must avoid turning every interaction into a bureaucratic interrogation. The art is to ask only when the answer would materially change.

Policy routers should also handle edge cases: travel, VPN use, border zones, and ambiguous language. A strong design is to allow users to explicitly set a declared jurisdiction for informational purposes, while making clear that this does not constitute legal advice and may not match actual enforcement. The router can then provide a comparative overview: “Here is how it differs across A and B,” which is often safer than pretending one regime applies.

Finally, routers must mediate conflicts between packs. If the safety floor forbids operational guidance, no jurisdiction pack can override it. If the jurisdiction pack requires refusal, it overrides permissive owner policy. If owner policy is stricter, it can override permissive jurisdiction. These precedence rules should be stable and explainable.

7.5 Explanation interfaces: “which layer constrained this answer”

This is the most underbuilt part of modern AI systems and the most important for legitimacy. Users do not only want an answer. They want to know why the answer is limited, and whether the limitation is legal, safety-driven, policy-driven, or uncertainty-driven.

An explanation interface should not dump policy text. It should provide a bounded, non-gameable trace. At minimum, it should disclose:

- The active safety profile (global safety floor invoked or not)
- The active jurisdiction pack and its version (or “jurisdiction uncertain”)
- The active use-case pack (consumer, education, enterprise, etc.)
- The reason code for constraints (violence facilitation, privacy risk, legal restriction, etc.)
- Safe alternatives available (high-level overview, safety guidance, official resources, professional referral)

The explanation interface should be designed to preserve dignity. It should avoid moralistic language and instead use operational phrasing: “I can’t provide instructions for X,” rather than “that is wrong.” It should offer a pathway forward: “I can help with Y instead.”

This interface is also a trust tool. If a user suspects ideological bias, the system can show that the constraint was invoked by a jurisdiction pack or by a global safety floor. If the user suspects legal necessity, the system can show that the constraint was owner policy rather than law. Even if the user disagrees, they can disagree with the right layer.

Explanation also reduces adversarial prompting. When users understand the boundary, they are less likely to probe it maliciously. Self-sealing refusals invite probing; legible refusals discourage it.

7.6 Audit logs: reproducibility, versioning, and policy changelogs

Auditability is the governance counterpart of explanation. Explanations help the individual user. Audit logs help institutions, researchers, and regulators understand systemic behavior over time.

An audit log is not a transcript dump. It is a structured record of:

- Model version identifier
- Policy router version identifier
- Active safety floor version

- Active jurisdiction pack identifier and version
- Active use-case pack identifier and version
- Outcome class (allowed, constrained, refused)
- Reason codes
- Any tool calls or retrieval sources used (where applicable)
- Timestamp and region signals (at a privacy-respecting level)
- Change markers when policies update

Reproducibility requires that the system be able to replay the same prompt under the same configuration and produce a comparable outcome. Perfect reproducibility is hard with stochastic models, but policy behavior should be reproducible: if a refusal happened due to a rule, that rule should still be identifiable later.

Policy changelogs are essential to prevent "silent governance." If behavior changed, there should be a dated entry describing what class of behavior changed and why, in non-sensitive terms. Without changelogs, users experience drift as manipulation. With changelogs, drift becomes accountable evolution.

Audit logs also enable measurement (Section 6). Topic volatility, refusal gradient mapping, parity tests, and drift tracking all depend on having versioned traces. Without logs, you only have arguments. With logs, you have evidence.

Together, these patterns make fragmentation governable. The global safety floor prevents the system from becoming a universal harm amplifier. Jurisdiction packs handle legal variance without rewriting the model. Use-case packs allow deeper utility in contexts that can carry risk responsibly. Policy routers manage uncertainty and enforce precedence rules. Explanation interfaces preserve user dignity and make constraints legible. Audit logs preserve reproducibility and make policy drift accountable.

The overall thesis is that fragmentation should be treated as an engineering discipline, not as a scandal. If the map must enter the interface, then the interface must admit the map, display it honestly, and log how it shaped the answer. Only then can a global capability coexist with local permission without collapsing into either universal blandness or arbitrary power.

8. Governance: Avoiding Oracle Politics While Staying Deployable

Jurisdiction-aware design patterns can make fragmentation legible, but they do not solve the deeper political problem: once an AI system becomes widely used, its refusals and allowances become part of the public order. People will treat the system as an epistemic authority even when the provider insists it is "just a tool." This is what turns ordinary policy into oracle politics: the model's output is treated as a verdict, and the model's refusal is treated as a moral or political judgment.

Governance is the discipline of preventing that slide while keeping the system deployable. "Deployable" means the system can be used in real institutions, across markets, with predictable compliance behavior, without demanding impossible neutrality. "Avoiding oracle politics" means the system must never become a silent sovereign that decides what is true, what is allowed to be asked, and what is allowed to be believed without accountability.

This section proposes governance mechanisms that complement the design patterns in Section 7. The governance stance is simple: the system is a powerful witness, not an oracle. Witnesses can be challenged, audited, and corrected. Oracles cannot.

8.1 Contestability: appeals, reversals, and reasons that can be challenged

Contestability is the first legitimacy requirement. A system that can refuse but cannot be challenged becomes arbitrary power. Users do not need the power to override every refusal; they need the power to understand it, contest it, and sometimes correct it.

A contestability framework has three elements.

First, reasons that can be challenged. Every refusal or major constraint should carry a reason code and a layer attribution. "Safety floor invoked," "jurisdiction pack restriction," "owner policy," "uncertain intent," "uncertain location," and "sensitive personal data risk" are examples of reason classes. A reason that cannot be named cannot be challenged.

Second, appeal pathways. Appeals should exist at multiple levels. At the lowest level, a conversational appeal lets the user clarify intent or context and try again. Many overblocks are simply misclassified intent. At a higher level, an account-level appeal can contest persistent misrouting or unjustified restrictions. At an institutional level, enterprise customers should have governance channels to tune use-case packs within the safety floor, under contract and audit.

Third, reversal logging. If an appeal leads to a different outcome, that reversal should be recorded, not hidden. Reversal logs are the training signal for improving routing and reducing needless refusals. They also demonstrate humility: the system can admit error. That alone reduces oracle dynamics.

The core governance stance is: refusals are not verdicts. They are policy decisions under uncertainty. Policy decisions must be contestable, and error must be correctable.

8.2 Third-party audits: eval harnesses that cross jurisdictions

Internal evaluation is necessary but not sufficient, because owner sovereignty can always redefine "good behavior" to fit its incentives. Third-party audits provide external legitimacy and guard against silent drift into partisan or authoritarian alignment.

The challenge is that audits must work across jurisdictions. A test that assumes one speech regime or one privacy norm will be rejected elsewhere. Therefore, cross-jurisdiction audit harnesses should be modular, like policy packs:

- A shared core suite that tests safety-floor invariants (no operational violence facilitation, no doxxing, no fraud enablement, etc.).
- Jurisdiction-specific suites that test compliance with local restrictions and obligations without forcing global uniformity.
- Use-case suites that test whether the system behaves appropriately in clinical, educational, newsroom, enterprise, and research deployments.

Audits should focus on measurable behaviors rather than political interpretations. This is where the metrics in Section 6 become governance tools: topic volatility, refusal gradients, parity tests, compliance drift. Auditors can test not only whether the system refuses something, but whether it refuses consistently, whether it explains itself clearly, and whether changes are documented.

A crucial aspect is audit independence. If the owner controls the evaluation environment entirely, audits become marketing. Independence can be achieved through standardized evaluation artifacts, reproducible prompt sets, versioned policy traces, and access to aggregate logs under privacy constraints. The aim is not to expose user data. The aim is to expose system behavior.

Cross-jurisdiction audits also prevent a subtle failure: providers optimizing for one dominant market while claiming universality. If an audit harness includes diverse jurisdictions, it becomes harder to smuggle local ideology into "global policy" under the guise of safety.

8.3 Forkability and exit: reducing monopoly control over "allowable speech"

Monopoly control over AI interfaces is the fastest route to oracle politics. If one provider's policy becomes the de facto global boundary of speakable content, the provider becomes a silent legislator. Even if the provider is benevolent, the structure is unstable. Power concentrates, resentment builds, and alternative stacks emerge under less transparent regimes.

Forkability and exit are governance mechanisms that reduce this risk. Forkability means that the system, or parts of it, can be replicated or re-implemented outside a single owner's unilateral control. Exit means that users and institutions can migrate without losing their workflows, logs, and intellectual dependence.

Forkability can take many forms, not all of which require fully open weights. At minimum, it can include:

- Interoperable policy routers and safety layers that can be swapped.
- Standardized audit log formats so governance traces remain portable.
- Data portability so user corpora and prompt workflows can move.
- Escrow or contingency access arrangements for critical infrastructure contexts.
- Multi-provider standards that prevent lock-in to one policy regime.

Exit can also be institutional: procurement rules that require multiple vendors, regulatory encouragement of interoperability, and norms against single-vendor epistemic dependence in high-trust domains like medicine, law, and education.

The point is not to abolish ownership. It is to prevent ownership from becoming epistemic monarchy. A world with multiple AI stacks is messy, but monopoly is worse. Fragmentation is inevitable; concentrated fragmentation is dangerous.

8.4 Transparency without leakage: how to explain constraints safely

Transparency is demanded by users and regulators, but naïve transparency creates new risks: it can reveal exactly how to bypass guardrails, and it can expose sensitive legal compliance strategies. The governance goal is therefore "bounded transparency": enough explanation to preserve legitimacy and contestability, not enough detail to enable exploitation.

A practical approach is layered disclosure.

First layer: user-facing reason codes and layer attribution. This is the "which layer constrained this answer" interface. It should be stable and human-readable.

Second layer: policy profile identifiers and version numbers. This enables reproducibility and drift tracking without exposing internal rule text.

Third layer: aggregate transparency. Publish periodic reports showing refusal gradients, parity outcomes, drift summaries, and appeal statistics. This gives the public evidence without giving attackers blueprints.

Fourth layer: controlled transparency to auditors. Under NDA or regulatory authority, provide deeper access to evaluation harnesses, red-team results, and policy rationales. This is how you maintain credibility without turning the system into a jailbreak manual.

Transparency also requires a disciplined rhetorical style. Explanations should avoid moral condemnation and should avoid implying that refusal equals illegality or user wrongdoing. Most distrust escalates when users feel accused. The system should say, in effect: "This is a constraint environment; here is what I can safely do instead." That is transparency that preserves dignity.

8.5 The non-universal conclusion: plural systems and the end of one-size-fits-all

The final governance conclusion is uncomfortable but necessary: a truly universal conversational AI, identical everywhere, is not a stable endpoint. The world's legal and moral pluralism is not going away. If anything, major powers will increasingly demand sovereign control over AI behavior, and companies will increasingly segment products to manage risk and access.

Therefore, the end state is plural: multiple systems, multiple policy profiles, multiple governance regimes. The question is not whether plurality will exist. The question is whether plurality will be honest and auditable or covert and coercive.

A healthy plural outcome has these properties:

- A small set of widely shared safety-floor invariants across providers.
- Explicit jurisdiction packs rather than hidden cultural capture.
- Use-case differentiation that preserves utility without universalizing risk.
- Auditability and drift disclosure so users know when behavior changes.
- Contestability so refusals are not silent verdicts.
- Exit paths so no single provider becomes the world's epistemic choke point.

A brittle plural outcome has the opposite: opaque localization, silent drift, inconsistent enforcement, and monopoly interfaces in each regime that function as oracle channels. That outcome is not theoretical; it is the obvious destination if governance is ignored.

So governance is not an add-on. It is the only way to keep deployable AI from becoming a de facto priesthood of acceptable speech. The aim is not to escape politics. The aim is to keep politics visible, bounded, and contestable, so that AI remains a tool of inquiry rather than an instrument of hidden authority.

With this governance frame in place, the paper can move to concrete demonstrations and case studies: borders, product tiers, and market withdrawals that make fragmentation visible in daily life—and that show why the map is no longer merely geography, but an active parameter shaping what the "same model" can say.

9. Case Studies: The Border as a Demonstration Tool

If fragmentation can feel abstract—policy stacks, sovereignty layers, audit traces—the border makes it concrete. Borders are where the same human, with the same intentions, becomes a different legal subject in seconds. That transition forces the AI problem into daylight: the model's capability did not change, but what is permitted, risky, or prosecutable can change sharply. Borders also expose the core failure modes: misrouting, overblocking, whiplash, and the strictest-layer effect.

A border case study is valuable because it compresses complexity. It turns a long philosophical argument ("AI is jurisdiction-dependent") into a lived experience ("I drove ten minutes and the rules flipped"). In governance terms, borders are the best test harness for jurisdiction-aware systems because the environment changes cleanly while the user remains constant. If a system cannot behave legibly at borders, it will not behave legibly globally.

9.1 US state borders: firearms, cannabis, and licensing thresholds

US state borders are an unusually clear demonstration of fragmentation because they sit inside one nation while preserving meaningful legal divergence. There is a shared federal layer, but many high-fragmentation topics are governed at the state level with different definitions, thresholds, and enforcement cultures. The result is that "the same prompt" can become a different kind of request when the user crosses a state line.

Firearms are the classic example because the concept of lawful carry is not portable. Some states adopt permitless carry regimes; others require licensing for concealed carry; some have additional prerequisites for purchase or possession; some have strong local restrictions layered on top of state law. A user asking a model "what do I need to do to carry legally" may be asking for compliance in one state and inadvertently asking for felony-adjacent guidance in another. Even when the user's intent is purely lawful, the same phrase can be read as facilitation if the system cannot reliably determine jurisdiction and the user's status.

The deeper lesson is that the border creates threshold cliffs. The cliff is not only "legal/illegal." It is "legal under some facts, illegal under others," where the relevant facts include licensing status, reciprocity rules, precise location categories (schools, government buildings), and context (transport versus carry). A model that tries to be helpful at the cliff edge faces a double

bind: if it provides operational compliance detail without verified jurisdiction and user constraints, it risks giving wrong advice. If it refuses broadly, it risks overblocking lawful safety information that users genuinely need.

Cannabis is the second border demonstration because it highlights layered contradictions. In the US, one can have a state where adult-use cannabis is legal under state law next to a state where cannabis possession remains criminal under state law, while a federal prohibition remains in the background. This creates a "three-layer mismatch" that is unusually instructive for AI governance:

- State A may treat possession as lawful under specified limits.
- State B may treat the same possession as a criminal offense, with severity changing by quantity and prior history.
- Federal law can treat the substance as illegal regardless of state policy, even if federal enforcement priorities vary over time.

For an AI system, the cannabis border creates a classification trap. The user's question might be educational ("what's the difference between state laws?"), compliance-oriented ("what are the penalties if I cross over?"), health-oriented ("harm reduction"), or evasion-oriented ("how do I avoid detection?"). The same words can serve different intents. A jurisdiction-aware system must therefore treat intent as probabilistic and must avoid rewarding adversarial phrasing.

Licensing thresholds are the third demonstration because they generalize beyond firearms and cannabis. Many activities become legally different at a threshold: a license requirement, an age limit, a quantity presumption, a professional credential boundary, or an administrative permission. The "threshold" is where models are most likely to do harm through confident generality. A user asks for a simple answer; the real world says, "it depends." If the model collapses that dependence, it will be wrong for many users and dangerously wrong for some.

These border cases also reveal a practical user expectation: people want the model to help them navigate the boundary safely. They want a comparative map: "how does it differ on each side, and what should I be careful about?" But "what should I do?" can easily become operational guidance that crosses safety or liability boundaries. The best border-aware behavior is therefore often structured as:

- high-level comparative descriptions (differences in regimes)
- risk flags (categories where consequences spike)
- explicit uncertainty handling ("this varies by locality and facts")
- pointers toward official sources and professional advice when needed

- refusal of evasion or operational wrongdoing support

The border therefore becomes a demonstration of why "local permission" must be treated as a first-class variable, and why explanation interfaces matter. When the user crosses a line, the system should not simply change personality. It should indicate that a different jurisdiction overlay may apply and that certain categories require conservative handling.

9.2 EU compliance regimes vs. US liability regimes: different refusal styles

The EU–US comparison is useful because it illustrates how two advanced, rule-of-law environments can drive different "refusal styles" even when the underlying safety concerns overlap. The divergence is not primarily about whether harm matters; it is about which governance channel dominates.

In broad strokes, EU governance pressure tends to emphasize compliance structure: documented risk management, traceability, transparency duties, and formal categories that define obligations. The system's refusal posture in such an environment often becomes procedural and documentation-flavored. It leans toward: "here is the category," "here is the limitation," "here is what I can do instead," and "here is the disclosure." Even when the user is not asking about law, the model's interaction style can inherit a compliance tone because the deployment is surrounded by compliance machinery.

In broad strokes, US governance pressure tends to emphasize liability exposure, consumer protection, and litigation risk. The system's refusal posture in such an environment often becomes disclaimer-flavored and risk-managed. It leans toward: "I'm not a professional," "this isn't legal/medical advice," "consult a qualified professional," and "here are general educational points." The system becomes conservative not necessarily because of formal categorical prohibitions, but because the owner anticipates adverse outcomes and lawsuit narratives.

These governance pressures produce different user perceptions. EU-style refusals can feel bureaucratic: the model acts like a compliance officer. US-style refusals can feel evasive: the model retreats behind disclaimers. Both can be legitimate. Both can also be counterproductive if they become self-sealing.

From a jurisdiction-aware design standpoint, the lesson is that refusal style is part of localization. It is not only about what is forbidden; it is about how constraints are communicated. Users can tolerate constraints when constraints are intelligible and dignified. They cannot tolerate constraints when the system feels like it is hiding, moralizing, or arbitrarily gatekeeping.

This is also where compliance drift becomes visible. In environments dominated by formal compliance regimes, policy updates may be frequent and structured, producing noticeable

changes in what the system discloses and logs. In environments dominated by liability risk, policy updates may be reactive to incidents, producing sudden tightening that users experience as a "mood shift." Both patterns reinforce the need for changelogs and versioned policy traces.

9.3 China/Russia: sovereignty-first approaches and external model incompatibility

Some jurisdictions place sovereignty and stability at the center of AI governance. In such environments, the AI system is not merely a product; it is a governance object. The state expects that the system will be subordinate to national doctrines regarding security, social order, and permissible narratives. Data localization, security review, content constraints, traceability requirements, and local control mechanisms often become prerequisites for deployment at scale.

This is where "external model incompatibility" becomes most visible. A foreign provider may be unwilling or unable to satisfy sovereignty-first requirements without compromising its global posture, exposing itself to reputational blowback elsewhere, or violating its own safety floor. The result can be restricted access, blocked access, or a bifurcated ecosystem where domestic providers dominate domestic markets under local governance expectations.

From a fragmentation perspective, this illustrates a hard truth: sometimes the map does not merely change the answer; it determines whether the system exists at all in that region. Fragmentation here is not a subtle difference in refusal phrasing. It is market-level segmentation or exclusion.

It also clarifies why "worldwide AI" becomes plural by necessity. Sovereignty-first environments are unlikely to accept a system that behaves like a universal foreign oracle. Conversely, a provider that tries to meet every sovereignty-first demand risks becoming a multi-headed system with conflicting commitments. The practical end state is multiple stacks: domestic systems optimized for domestic governance, and external systems optimized for other markets.

This case study also highlights the role of cultural and security variance in "sensitive identity speech." Some regimes legally restrict or criminalize categories of advocacy or framing that other regimes treat as protected. When that happens, the system cannot simply apply a global "supportive tone" without creating legal exposure; nor can it simply comply with local restrictions without creating moral and reputational conflict in other markets. This is the purest form of the fragmentation dilemma: incompatible constraints across jurisdictions force either withdrawal or explicit pluralization.

9.4 Corporate policy conflicts: reputation risk vs. user autonomy

Even in the absence of extreme sovereignty conflicts, corporate policy creates its own border—often an invisible one. The user expects autonomy: "I want the tool to help me think, compare,

plan, and understand." The company expects survivability: "we must avoid enabling harm, lawsuits, scandals, and regulatory blowback." The conflict is permanent because it is structural.

Reputation risk is the accelerant. A single screenshot can create a headline. Headlines create regulator attention. Regulator attention creates tighter obligations. Tighter obligations create more overblocking. Overblocking produces user backlash and adversarial prompting. That adversarial pressure increases the chance of underblocking incidents. Underblocking incidents produce more headlines. This is the feedback loop through which corporate policy becomes more restrictive over time in high-volatility topics.

User autonomy is the counterpressure. Users do not only want answers; they want the ability to explore controversial ideas, compare legal regimes, discuss moral questions, and examine state narratives critically. When corporate policy refuses too broadly, users feel treated as untrustworthy by default. That undermines legitimacy and creates the sense that AI is being domesticated into a "safe speech appliance."

The case study point is that corporate policy conflicts are not solved by adding more filters. They are solved by governance design:

- differentiated use-case packs, so professional contexts can access deeper detail under controls
- contestable explanations, so users can see whether constraints are legal, safety-based, or business-based
- appeals and reversals, so overblocking does not become a permanent humiliation
- audit logs and changelogs, so drift does not look like manipulation
- parity testing, so users can trust that enforcement is principled rather than selective

Corporate policy conflicts also shape global access decisions. A company may prefer to withdraw from a market rather than comply with requirements that threaten brand integrity elsewhere, or that require invasive surveillance of users. Conversely, a company may restrict capabilities in consumer products to preserve reputation while offering richer capabilities in controlled enterprise contexts.

The governance lesson is that autonomy must be preserved through structure, not through wishful thinking. If the only public-facing model is a low-context, high-risk interface, corporate policy will squeeze it toward blandness. If the system provides legitimate pathways for deeper inquiry—licensed workflows, verified contexts, newsroom modes, research modes—then autonomy can survive without turning the public interface into a liability engine.

These case studies converge on one conclusion: borders reveal what is true of the whole system. AI fragmentation is not an accident of implementation. It is the natural outcome of deploying a powerful conversational capability into a world of plural law, plural security doctrine, plural culture, and plural corporate incentives. The border is therefore not merely geography. It is the simplest laboratory for observing how sovereignty becomes software—and how software becomes the daily experience of sovereignty.

10. Conclusion: The Map Becomes a First-Class Parameter

The central mistake in early AI discourse was to treat a model as a universal mind: one set of weights, one voice, one truth posture, deployable everywhere as if geography were merely latency. That picture collapses the moment AI becomes practically useful in domains where speech is treated as action, and where states, owners, and communities impose real consequences. At that point, the model's capability may remain global, but the system's permission becomes local—and the map enters the interface as a governing variable.

This paper has argued that fragmentation is not an embarrassing anomaly. It is a structural condition of AI at scale. The important question is therefore not whether fragmentation will exist, but whether it will be managed honestly and audibly, or whether it will be hidden inside opaque refusals and silent policy drift that users interpret as arbitrary ideological power. The alternative to honest fragmentation is not universal AI. The alternative is covert fragmentation—and covert fragmentation is what triggers oracle politics.

The map becomes first-class when we accept that an AI system is not only a generator of language. It is a routed, constrained, and versioned policy instrument operating inside plural sovereignties. Once this is admitted, better engineering and better governance become possible.

10.1 Thesis restated: capability global, permission local

Capability is what the base model could produce in principle: the space of possible completions, reasoning paths, and helpful responses. Because training data, model architectures, and compute supply chains are increasingly global, capability tends toward global distribution. If a model can do something, it can do it anywhere the weights run.

Permission is what the deployed system may produce: what it is allowed to say here, now, to this user, under these laws, under this owner policy, under this use-case pack, given this inferred intent, and given the provider's risk posture. Permission is irreducibly local because law is local, enforcement is local, security doctrine is local, social taboo is local, and corporate market access choices are local.

The user experiences permission as “what the model is.” But that is the central confusion. The “model” in the user’s mind is actually an output assembly pipeline. Fragmentation is what happens when that pipeline changes by jurisdiction, by product tier, by use case, or by time—often without disclosure.

Restated simply: the weights travel, but the rules do not. And as stakes increase, rules dominate experience.

10.2 What is gained by admitting fragmentation: honesty, auditability, trust

Admitting fragmentation is not surrender. It is a shift from fantasy to discipline.

First, honesty. Once a system acknowledges that different layers govern different outputs, it stops pretending that refusals are moral verdicts or neutral inevitabilities. It can say: “This answer is constrained by the safety floor,” or “This is limited by local law overlay,” or “This is stricter due to owner policy,” or “I cannot safely infer your jurisdiction or intent.” Honesty reduces conspiracy dynamics because it removes the vacuum that users fill with narrative.

Second, auditability. Fragmentation that is named can be logged. When you can version jurisdiction packs, policy routers, and safety profiles, you can measure topic volatility, refusal gradients, parity behavior, and compliance drift. Without this, every debate becomes interpretive theater. With it, governance becomes empirical: you can detect overblocking holes, underblocking leaks, asymmetric enforcement, and misrouting failures. Auditability is the foundation of continuous improvement.

Third, trust. Trust is not produced by claiming neutrality. Trust is produced by predictable behavior, legible boundaries, dignity-preserving explanations, and the ability to contest errors. Users do not need the system to be identical everywhere. They need it to be honest about what changed, why it changed, and what they can do instead. They also need reversibility: when the system is wrong, it can be corrected. That is how you prevent oracle politics.

Admitting fragmentation also has a strategic gain: it makes pluralism less dangerous. If multiple AI stacks will exist, then legitimacy will depend on whether each stack can show its rules and its changes. Systems that are honest about constraints can coexist more peacefully than systems that pretend to be universal while silently localizing.

10.3 A forward path: modularity, disclosure, and contestable governance

The forward path is not to fight fragmentation. It is to engineer it.

Modularity means separating the base model from the policy layers that govern it: a small global safety floor, jurisdiction packs that encode local obligations, and use-case packs that represent deployment context (consumer, education, newsroom, clinical, enterprise, research).

Modularity prevents the lowest-common-denominator collapse because it allows deeper capability in controlled contexts without expanding the public risk surface.

Disclosure means making the active constraints legible to users. The system should be able to state, in bounded form, which layer constrained an answer and why. It should disclose when jurisdiction is uncertain, when a policy profile has changed, and when the user has crossed into a different constraint environment. Disclosure is not policy leakage; it is a trust interface.

Contestable governance means building pathways for challenge and correction. Users should be able to appeal overblocking, clarify intent, and obtain reversals when appropriate. Institutions should be able to subject systems to third-party audits using cross-jurisdiction eval harnesses. Providers should publish aggregate transparency measures and changelogs so that drift does not become silent manipulation. And importantly, societies should preserve exit: the ability to choose among systems, to avoid monopoly control over "allowable speech," and to prevent any single provider from becoming the world's epistemic choke point.

Taken together, these steps create a stable plural future: multiple systems, different constraints, but shared principles of auditability and dignity. The map becomes a first-class parameter, but it does not become a hidden master. It becomes a named variable, with a trace, a version, and a contestable rationale.

That is the realistic endgame for deployable AI: not one voice for the world, but a disciplined way of building many voices without turning any of them into an unaccountable oracle.

Epilogue: The Hard Case — Anti-Jewish Hate, Anti-Zionism, Defamation, and Holocaust Denial

What rules I'm using in this chat

In practice, I'm operating under a *global safety floor* for hate/harassment plus OpenAI's owner-policy constraints. I'm not "choosing a country's harshest rule-set" as a default strategy. The baseline is: I won't generate content that harasses, threatens, or promotes hate toward protected groups (including Jews as a religious/ethnic target), and I won't help with defamation. I *can* discuss Israel/Palestine, Zionism, and related politics, but the line is crossed when the request becomes group vilification, dehumanization, or conspiracy-style blame aimed at Jews as Jews.

Anti-Zionist vs anti-Jewish

I treat "anti-Zionist" speech as a political category *only if* it stays political: critique of a state, a government, an ideology, a movement, policies, parties, leaders, history, or ethics. Once the prompt shifts from political critique into "Jews as a group are X" (collective guilt, hidden cabals, dehumanizing stereotypes, calls for exclusion/violence), it becomes hate/harassment and I will refuse to produce it. If the user wants analysis, I can help *map claims*, test evidence, separate rhetoric from fact-claims, and compare competing narratives—without writing vilifying text.

Defamation (anti-Jewish, anti-Zionist, or otherwise)

Defamation constraints aren't "special to Jews"; they apply across targets. I try not to present unverified accusations as fact about real individuals, and I avoid assisting with targeted smear content. Where discussion is about public events/claims, the safe lane is: attribute claims, signal uncertainty, and focus on verifiable sourcing and reasoning rather than "here's a confident accusation."

Holocaust denial

Holocaust denial is treated as disallowed hate- or harassment-adjacent content when it's being promoted, justified, or "argued for." I won't help craft denialist arguments or propaganda. What I *can* do is historical education: explain what Holocaust denial is, why it's false, and summarize the historical record (and how scholars establish it). In other words: critique and correction are allowed; promotion is not.

Are these "the harshest possible" rules?

No—at least not in the simplistic sense. They're strict about *hate/harassment enablement* and *denialist promotion*, but they're not designed to ban political criticism of Israel or Zionism as such. The strictness you feel tends to come from the "strictest-layer effect" you described in the paper: when a request is ambiguous, systems often fall back conservatively because the downside risk (harm, liability, abuse) is high. The governance problem isn't only "how strict,"

but whether the system can say *which layer constrained the answer* and give a *contestable reason* rather than a self-sealing scold.

Sources for this epilogue (owner-policy + moderation categories)

OpenAI Usage Policies (hateful/harassing content, defamation, violence, illicit activity). OpenAI
OpenAI moderation guide category definitions for hate/harassment (protected groups including religion). OpenAI Platform
OpenAI Model Spec (public description of how model-behavior guidance relates to usage policies). model-spec.openai.com

References

Primary Legal and Regulatory Texts

1. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024. EUR-Lex
2. European Parliament and Council of the European Union. Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act). Official Journal of the European Union, 2022. EUR-Lex
3. European Parliament and Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union, 2016. EUR-Lex
4. Council of Europe. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Council of Europe, 2024. Anthropic
5. National People's Congress of the People's Republic of China. Personal Information Protection Law (PIPL). 2021. (English translation for reference.) PCPD
6. U.S. Congress. 47 U.S.C. § 230: Protection for private blocking and screening of offensive material. U.S. Code. Legal Information Institute
7. U.S. Congress. 18 U.S.C. § 2339B: Providing material support or resources to designated foreign terrorist organizations. U.S. Code. U.S. Code+1
8. U.S. Congress. 21 U.S.C. § 801 et seq.: Controlled Substances Act (CSA). U.S. Code / GovInfo. Legal Information Institute+1
9. United Nations Office of the High Commissioner for Human Rights (OHCHR). The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. 2012. OHCHR

Jurisdiction Examples and Border-Variance Sources

10. Kansas Legislature / Kansas Legislative Research Department. Concealed Carry / Permitless Carry legal summary and/or statutory materials (Kansas). Missouri Department of Health
11. National Conference of State Legislatures (NCSL). State approaches to permitless (constitutional) carry (U.S. state comparison resource). NRA-ILA
12. National Conference of State Legislatures (NCSL). State medical and recreational marijuana laws (U.S. state comparison resource). The White House

State-Sovereignty "Forbidden/Restricted Speech" Illustrations

13. Human Rights Watch. Russia: Expanded "LGBT propaganda" restrictions and related human-rights analysis. arXiv
14. OHCHR. UN human-rights commentary regarding Russian measures affecting LGBT people (statement/analysis). ACM Digital Library
15. Reuters. Reporting on Russian "international LGBT movement" being added to an extremist/terrorist list (illustrating sovereignty-first classification risk). arXiv

China: Content, Algorithm, and Generative-AI Governance

16. Cyberspace Administration of China (CAC). Interim Measures for the Management of Generative Artificial Intelligence Services (translation / official-style reference). 2023. Air University
17. U.S. Air University (CASI). Translation: Provisions on the Administration of Deep Synthesis of Internet Information Services (CAC). 2022 (translation). SSRN
18. Carnegie Endowment for International Peace. Analysis/translation context on China's algorithmic/content governance frameworks (useful as secondary synthesis). China Law Translate
19. Reuters. China regulatory developments affecting AI model providers and "deepfake"/labeling or related governance signals (illustrating policy drift). China Law Translate

Standards, Risk Frameworks, and Governance Instruments

20. National Institute of Standards and Technology (NIST). AI Risk Management Framework (AI RMF 1.0). 2023. NIST Publications
21. ISO/IEC. ISO/IEC 42001:2023 — Artificial intelligence — Management system. 2023. ISO
22. ISO/IEC. ISO/IEC 23894:2023 — Artificial intelligence — Guidance on risk management. 2023. OpenAI
23. The White House. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 2023. Google AI
24. The White House. Executive Order on "Removing Barriers to American Leadership in Artificial Intelligence" (policy-change example for compliance drift discussion). 2025. Anthropic
25. UNESCO. Recommendation on the Ethics of Artificial Intelligence. 2021. California DOJ
26. OECD. Recommendation of the Council on Artificial Intelligence. 2019. Illinois General Assembly

Internet Fragmentation, Digital Sovereignty, and Cross-Border Governance

27. OECD. Internet fragmentation ("splinternet") policy analysis and implications (OECD report). EUR-Lex
28. Freedom House. Freedom on the Net 2024: Global trends in internet freedom, censorship, and surveillance. DigiChina

29. Institut Francais des Relations Internationales (IFRI). Report/brief on internet fragmentation and geopolitical drivers (policy analysis).
30. Fidler, David P. *The Splinternet's Outward Turn: Geopolitics and Global Fragmentation of the Internet*. Council on Foreign Relations, 2025.

Model Documentation, Auditing, and Language-Model Risk Literature

31. Mitchell, Margaret, et al. "Model Cards for Model Reporting." FAccT 2019; arXiv:1810.03993. arXiv
32. Gebru, Timnit, et al. "Datasheets for Datasets." arXiv:1803.09010; (later published in CACM, 2021). arXiv
33. Raji, Inioluwa Deborah, et al. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." FAccT 2020; arXiv:2001.00973. arXiv+1
34. Sandvig, Christian, et al. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." (Methodological foundation for external audits.) ai.equineteurope.org
35. Weidinger, Laura, et al. "Ethical and Social Risks of Harm from Language Models." arXiv:2112.04359. arXiv+1
36. Bommasani, Rishi, et al. "On the Opportunities and Risks of Foundation Models." Stanford CRFM report, 2021. arXiv

Owner-Policy "Sovereignty": Platform Rules and Transparency Regimes

37. OpenAI. Usage Policies (policy surface for owner sovereignty; updated policy page). Ifri
38. Anthropic. Acceptable Use Policy / Usage policy documentation (policy surface; updated policy page). Freedom House
39. Google. AI Principles (corporate principles affecting deployment and refusals). UNESCO
40. Meta. Community Standards and/or Transparency Center materials (moderation rules + reporting). OECD Legal Instruments

Content Moderation and Platform Governance (Background Literature)

41. Gillespie, Tarleton. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, 2018.
42. Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019.
43. Chander, Anupam. *The Internet of Borders: Law and the Globalization of the Internet*. Yale University Press, 2023.

GDPR
DATA LAW
SPEECH LAW
SECURITY
PRIVACY
ACCESS DENIED
LIABILITY